

## Statistical Methods for Longitudinal Research

### Autumn 2020 Remote Asynchronous Instruction

David Rogosa Sequoia 224, [rag@stanford.edu](mailto:rag@stanford.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](#)

The positive results continued beyond the end of the treatment period. The mean reduction on the depression scale in the treatment arm remained statistically superior to that of the placebo group two weeks after dosing stopped.

### Questions

Consider the remission outcome (secondary) at day 15 (after 14 days of dose).

part a. For these time1-time2 dichotomous data (remission or not), explain what I did below to approximate the results reported by SAGE.

part b. In week 4 (time1-time2 data) materials we introduced some more advanced capabilities for time1-time2 dichotomous data, such as `mcnemar.test` from base R and `diffpropci.mp` from package `PropCIs`. Comment on the applicability of those functions to the remission study and whether those are preferable here to the basic analysis in part a.

```
-----
> sage2 = matrix(c(29, 10, 16, 34), nr=2) # remission counts for the two groups
> sage2
     [,1] [,2]
[1,]   29  16
[2,]   10  34
> prop.test(sage2)
      2-sample test for equality of proportions with continuity correction
data:  sage2
X-squared = 14.078, df = 1, p-value = 0.0001754
alternative hypothesis: two.sided
95 percent confidence interval:
 0.2079003 0.6264431
sample estimates:
   prop 1    prop 2 
0.6444444 0.2272727

> chisq.test(sage2)
      Pearson's Chi-squared test with Yates' continuity correction
data:  sage2
X-squared = 14.078, df = 1, p-value = 0.0001754
-----
```

part c. Consider the primary outcome, change in depression score (HAM-D).

In weeks 4 and 5 we conducted analysis of time1- time2 (and multiwave) outcome data for comparisons of experimental groups. For the SAGE study pretend we have long form data, with time coded 0 for baseline and 1 for Day 15 endpoint, and outcome HAM-D score at the timepoints (0,1) and group indicating T/P. So we have 178 rows, and columns HAM-D group time subj.

If we fit the model in R syntax

```
sage1mer = lmer(outcome ~ time + time:group + (time|subj), data = sage, control = lmerControl(check.nobs.vs.nRE = "warning"),
from the information you have, give the point estimate for the fixed effects, time and time:group .
```

Write out the level 1, level 2 model corresponding to the combined model in the lmer statement.

## Week 5. Experimental Protocols and Comparing Group Growth

From the *Longitudinal in the news* archive

### Time 1, Time 2 Experiments

[When Adolescents Give Up Pot, Their Cognition Quickly Improves](#)

From 2017. Stents? [A Controversial Experiment Upends The Conventional Wisdom On Heart Stents](#) Publication: [Percutaneous coronary intervention in stable angina \(ORBITA\): a double-blind, randomised controlled trial](#) The Lancet.

### Crossover Designs in the news

1. Does nutrition science know anything? [Is white or whole wheat bread 'healthier?' Depends on the person](#) Publication: [Bread Affects Clinical Parameters and Induces Gut Microbiome-Associated Personal Glycemic Responses](#) Cell Metabolism, Korem et al DOI: 10.1016/j.cmet.2017.05.002
2. This time with 3 conditions [For Exercise, Nothing Like the Great Outdoors](#) Publication: Niedermeier M, Einwanger J, Hartl A, Kopp M (2017) [Affective responses in mountain hiking-- randomized crossover trial focusing on differences between indoor and outdoor activity](#). PLoS ONE 12(5): e0177719. <https://doi.org/10.1371/journal.pone.0177719>
3. One thing at a time. [Why listening to a podcast while running could harm performance](#). Publication: [A trade-off between cognitive and physical performance, with relative preservation of brain function](#) Scientific Reports 7, Article number: 13709 (2017) nature.com.
4. Another crossover design (from Stat266). RCT (cross-over design). [Damn right! The secret of success is swearing: How shouting four letter words can help make you stronger](#) [Swearing can help you boost your physical performance](#) [The full power of swearing is starting to be discovered](#). Publication: Stephens, R, Spierer, DK and Katehis, E (2018) Effect of swearing on strength and power performance. Psychology of Sport and Exercise, 35. pp. 111-117. ISSN 1469-0292 .

### Lecture Topics

1. Cross-over designs (usually time-1, time-2). [Laird-Ware text slides](#) (pdf pages 135-150). [Crossover design data from slide 137](#), [anova for crossover design ex](#) [ascii version](#), [anova for crossover design ex](#)

R-resources for crossover designs. package [Crossover](#) package [crossdes](#) see Rnews Vol. 5/2, November 2005  
also see slides 5-14 [Repeated Measures Design Mark Conaway](#).

2. [Multi-wave growth in measured outcome](#), experimental protocol. Example: Effect of transitional probability (TP) on novel word learning from Mirman text Ch3 (uses orthogonal polynomials). Data from Week3 Review Question 4.

3. [Group Comparisons for Longitudinal Experimental Designs](#). Group growth and Experimental comparisons for count and dichotomous outcomes (examples From HSAUR 2ndEd, Ch.13).

Link functions for generalized linear mixed models (GLMMs), [Bates slides](#) (pdf pages 11-18)

Note: glmer supercedes lmer

A Handbook of Statistical Analyses Using R, Second Edition Torsten Hothorn and Brian S. Everitt Chapman and Hall/CRC 2009. [Analysing Longitudinal Data II -- Generalised Estimation Equations and Linear Mixed Effect Models: Treating Respiratory Illness and Epileptic Seizures](#) (Stanford access)

Data sets etc [Package 'HSAUR2'](#) August 2014, Title A Handbook of Statistical Analyses Using R (2nd Edition)

Note: epilepsy available in all vintages of HSAUR

**A.** [Analysis of Count data.](#) Epilepsy example, group comparisons, collection of individual trajectories. HSAUR chap 13 [Rogosa R-session using gee and lmer](#) [class handout](#)

Recap Group Comparisons, Epilepsy example. [Comparison of lmer models](#)

For SAS (and GEE) fans [another analysis](#)

**B.** [Binary Response, dichotomous outcomes.](#) Respiratory Illness Data from HSAUR package: [Data and description](#) also at the ALA (Laird-Ware) site [Rogosa R-session using lmer](#) [class handout](#)

#### 4. [Study Design: Power Calculations](#) for Longitudinal Group Comparisons.

R-package [longpower](#) Vignettes found by "browseVignettes(package = "longpower")". Functions in [MBESS](#) package--[ss.power.pcm](#).

R-package: [powerlmm](#)

Background pubs: [Power for linear models of longitudinal data](#) with applications to Alzheimer's Disease Phase II study design Michael C. Donohue, Steven D. Edland, Anthony C. Gamst

[Sample Size Planning for Longitudinal Models:](#) Accuracy in Parameter Estimation for Polynomial Change Parameters Ken Kelley Notre Dame Joseph R. Rausch *Psychological Methods* 2011

basic R analogues, `power.t.test` `power.anova.test`

#### 5. Missing Data Concerns.

Nontechnical overviews:

Phil Lavori et al. *Psychiatric Annals*, Volume 38, Issue 12, December 2008 Missing Data in Longitudinal Clinical Trials, [Part A](#) [Part B](#)

Robin Henderson, [Missing Data in Longitudinal Studies](#) pdf pages 89-93

*We probably won't get to this (usually defer to DW)*

Missing data wide-form imputation: mice multiple regression example, nhanes data in package mice [R-session using mice package](#)

New package: hmi: hierarchical multiple imputation, [vignette](#)

Vignette from merTools package (Stat196): [Analyzing Imputed Data with Multilevel Models and merTools](#)

[Rogosa R-session for vignette](#)

*Additional resources.*

Technical review: [Missing data methods in longitudinal studies: a review](#) Joseph G. Ibrahim corresponding author and Geert Molenberghs

More on Missing data and imputation, including mice week 10 topic. Flexible Imputation of Missing Data. Stef van Buuren Chapman and Hall/CRC 2012. [Chapter 9, Longitudinal Data](#) Sec 3.8 Multilevel data. He is the originator of mice

Multiple Imputation. van Buuren S and Groothuis-Oudshoorn K (2011). [mice: Multivariate Imputation by Chained Equations in R](#). *Journal of Statistical Software*, 45(3), 1-67. see also [multiple imputation online](#) Flexible Imputation of Missing Data. Stef van Buuren Chapman and Hall/CRC 2012. [Chapter 9, Longitudinal Data](#) Sec 3.8 Multilevel data. He is the originator of mice [book extras](#) R resources. [Multivariate Analysis Task View](#), *Missing data* section, esp packages mice see also [multiple imputation online](#)

[CHAPTER 17 Incomplete data: Introduction and overview.](#) *Longitudinal Data Analysis* Edited by Geert Verbeke, Marie Davidian, Garrett Fitzmaurice, and Geert Molenberghs Chapman and Hall/CRC 2008. Also CHAPTER 21 Multiple imputation Michael G. Kenward and James R. Carpenter and CHAPTER 22 Sensitivity analysis for incomplete data. [online supplement for LDA book](#). van Buuren S (2010). [Multiple Imputation of Multilevel Data](#). In JJ Hox, K Roberts (eds.), *The Handbook of Advanced Multilevel Analysis*, chapter 10, pp. 173-196. Routledge, Milton Park, UK [Handling drop-out in longitudinal studies](#) (pages 1455-1497) Joseph W. Hogan, Jason Roy and Christina Korkontzelou, *Statistics in Medicine* 15 May 2004 Volume 23, Issue 9. (SAS implementations)

Bayesian approach. [Missing Data in Longitudinal Studies. Strategies for Bayesian Modeling and Sensitivity Analysis](#) Joseph W. Hogan and Michael J. Daniels Chapman and Hall/CRC 2008 Ch 5 Missing Data Mechanisms and Longitudinal Data [Corresponding talk](#), A Brief Tour of Missing Data in Longitudinal Studies Mike Daniels

Overview and applications paper: [Assessing missing data assumptions in longitudinal studies: an example using a smoking cessation trial](#) Xiaowei Yanga, Steven Shoptawb. *Drug and Alcohol Dependence* Volume 77, Issue 3, 7 March 2005, Pages 213-225

R resources. [Multivariate Analysis Task View](#), *Missing data* section, esp packages mice and mi R-package pan Multiple imputation for multivariate panel or clustered data. [Schafer tech report](#) Schafer talk: [Missing Data in Longitudinal Studies: A Review](#) [Efficient ways to impute incomplete panel data](#). Kristian Kleinke A. Mark Stemmler A. Jost Reinecke A. Friedrich Losel AStA Adv Stat Anal (2011) 95:351-373 DOI 10.1007/s10182-011-0179-9

#### WEEK 5 Review Questions

1. Power (sample size) calculations for experimental group comparisons.

a. Longpower package (vignette). Reconstruct the sample size calculation for the Alzheimer's disease trial (7 waves) on p.4 of the vignette.

b. MBESS package. Recreate the sample size calculation for width of confidence interval for differential growth using `ss.aape.pcm` function in the example used in Kelley and Rausch appendix (and MBESS manual)

[Solution for Review Question 1](#)

2. Revisit Respiration example.

a. try to do lmlist on these data to get odds(good) for each of the each 111 subjects. Investigate effectiveness of treatment.

b Use lmer analyses to compare treatment and placebo. Obtain a confidence interval for effectiveness of treatment. Investigate gender differences in response to the intervention (i.e. the treatment)

c. Extend the lmer model in part b by adding the age and baseline measurements to the level 2 model. Compare with part b results.

[Solution for Review Question 2](#)

3. Revisit Epilepsy example.

To supplement the longitudinal texts (HSAUR, ALA etc) full model for the epilepsy data, let's try to build up the analysis from basic description comparing placebo vs drug up through some basic some basic glmer models.

A somewhat similar effort was made in the second class posting "Recap group comparisons (epcomp)" linked above. In this exercise treat period as a time measurement (1,2,3,4) rather than an ordered factor.

How many subjects in placebo and drug groups? Use lmlist to obtain slopes and intercepts for fits of time trends to seizures for each subject and compare drug and placebo groups.

Fit and compare glmer models with treatment as the only level 2 predictor (for intercept) without and with a time trend. Compare. Add the baseline to the glmer models above (in level 2 model for intercept; is effect of the drug significant (use confint)? Does adding age help this model?

#### [Solution for Review Question 3](#)

4. Extensions for the epilepsy example: residuals, diagnostics  
Revisit and extend analysis of ep3 model in RQ3 and lecture.

Obtain confidence interval for effect of drug, look at residual plots and identify any anomalous individuals, try out `merTools` package for additional plots and evaluations of uncertainty.

#### [Solution for Review Question 4](#)

5. Revisit cross-over design, class example, Lecture item 1. The class example used repeated measures analysis of variance for estimation the effect of the drug in the dialysis example, (I messed up the medical context in class). Repeat that analysis using lmer and show identical results to class example analysis. Also examine the effectiveness, increase in precision, resulting from each subject functioning as their own comparison, rather than having two separate (randomly assigned) treatment and control groups.

#### [Solution for Review Question 5](#)

### WEEK 5 Exercises

1. We use a subset of the Baumann data from the `car` package, which I was nice enough to put in longform at

<http://rogosateaching.com/stat222/readlongdat>.

These data are from a study of reading from Purdue. We use the data to compare two methods: Basal, traditional method of teaching; DRTA, an innovative method; coded 1 and 2 respectively in the data. Random assignment placed twenty-two students in each group; reading test measures were obtained pre and post instruction.

The Directed Reading Thinking Activity (DRTA) is a strategy that guides students in asking questions about a text, making predictions, and then reading to confirm or refute their predictions. The DRTA process encourages students to be active and thoughtful readers, enhancing their comprehension. Use descriptive and inferential statistical methods to assess the relative efficacy DRTA method.

2. Treatment of Lead Exposed Children (TLC) Trial. Data (wide form) and description: [data here](#)

Start out by just using the subset of the longitudinal data Lead Level Week 0 and Week 6. Carry out the repeated measures anova for the relative effectiveness of chelation treatment with succimer or placebo (A,P). Show the three equivalences in the Brogan-Kutner paper between the repeated measures anova results and simple t-tests for these data. Next compare with a lmer fit following the B-K class example (posted). Finally use all 4 longitudinal measures (weeks 0,1,4,6) for a Active vs Placebo comparison using lmer. Compare with the results that use only 2 observations.

3. Crossover Design. The dataset consists of safety data from a crossover trial on the disease cerebrovascular deficiency. The response variable is not a trial endpoint but rather a potential side effect. In this two-period crossover trial, comparing the effects of active drug to placebo, 67 patients were randomly allocated to the two treatment sequences, with 34 patients receiving placebo followed by active treatment, and 33 patients receiving active treatment followed by placebo. The response variable is binary, indicating whether an electrocardiogram (ECG) was abnormal ( $Y=1$ ) or normal ( $Y=0$ ). Each patient has a bivariate binary response vector.

Data set is available at <http://www.hsph.harvard.edu/fitzmaur/ala/ecg.txt> (needs to be cut-and-paste into editor). Carry out the basic analysis of variance for this crossover design following week 5 Lecture topic 2. You may want to use glm to take into account the binary outcome. Does the treatment increase the probability of abnormal ECG? Give a point estimate and significance test for the treatment effect.

4. Data on Amenorrhea from Clinical Trial of Contracepting Women. Source: Table 1 (page 168) of Machin et al. (1988). Reference: Machin D, Farley T, Busca B, Campbell M and d'Arcangues C. (1988). Assessing changes in vaginal bleeding patterns in contracepting women. *Contraception*, 38, 165-179.

Data [in long form](#) and [a wide-form version](#)

Description: The data are from a longitudinal clinical trial of contracepting women. In this trial women received an injection of either 100 mg or 150 mg of depot-medroxyprogesterone acetate (DMPA) on the day of randomization and three additional injections at 90-day intervals. There was a final follow-up visit 90 days after the fourth injection, i.e., one year after the first injection.

Throughout the study each woman completed a menstrual diary that recorded any vaginal bleeding pattern disturbances. The diary data were used to determine whether a woman experienced amenorrhea, the absence of menstrual bleeding for a specified number of days. A total of 1151 women completed the menstrual diaries and the diary data were used to generate a binary sequence for each woman according to whether or not she had experienced amenorrhea in the four successive three month intervals.

In clinical trials of modern hormonal contraceptives, pregnancy is exceedingly rare (and would be regarded as a failure of the contraceptive method), and is not the main outcome of interest in this study. Instead, the outcome of interest is a binary response indicating whether a woman experienced amenorrhea in the four successive three month intervals. A feature of this clinical trial is that there was substantial dropout. More than one third of the women dropped out before the completion of the trial. In the linked data, missing data are designated by "." [note: in the week 6 terminology consider the dropouts to be *missing at random*, not necessarily a correct assumption.]

The purpose of this analysis is to assess the influence of dosage on the risk of amenorrhea and any individual differences in the risk of amenorrhea.

Show your model for these data and the results. Provide significance tests and/or interval estimates for the odds of amenorrhea as a function of dose.

Display and interpret individual differences in response by showing the random effects within each experimental group.

5. Chick Data, *finale*. One more use of the chick data (week 3, problem 2; week 1 class lecture). Use the data for all 4 Diets to construct a nlmer model that allows asymptotes to differ across the four diets. Do the diets produce significantly different results? Which diet produces the heaviest 'mature' chick weight?

6. Missing Data. Wide-form longitudinal data

Artificial data example from week 2 RQ3 and Week 4 Lecture item 4 (used in Myths examples to illustrate time-1, time-2 data analysis) [Two part artificial data example](#). The top frame (the  $X_i$ 's) is 40 subjects each with three equally spaced time observations (here in wide form). For these these perfectly measured " $X_i$ " measurements each subject's observation fall on a straight-line.

a. Use data set [W6probla](#), for which about 15% of the observations have been made missing. Use these data (with lm) to recreate the multiple regression demonstration in Week 4 lecture, part 4: "Correlates and predictors of change: time-1, time-2 data". Compare with the results for the full data on 40 subjects. What does lm do with missing data?

b. Repeat part a with data set [W6problb](#). Can you find any reason to doubt a "missing at random" assumption for this data set?

Note: if we don't get to it in Week 5, then in Week 10 (DW) we will demonstrate multiple imputation procedures (`mice`) for wide-form data, at least.

---

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

### Lecture Topics



## One Month of Cannabis Abstinence in Adolescents and Young Adults Is Associated With Improved Memory

Randi Melissa Schuster, PhD; Jodi Gilman, PhD; David Schoenfeld, PhD; John Evenden, PhD; Maya Hareli, BA; Christine Ulysse, MS; Emily Nip, BA; Ailish Hanly, BA; Haiyue Zhang, MS; and A. Eden Evins, MD, MPH

**Objective:** Associations between adolescent cannabis use and poor neurocognitive functioning have been reported from cross-sectional studies that cannot determine causality. Prospective designs can assess whether extended cannabis abstinence has a beneficial effect on cognition.

**Methods:** Eighty-eight adolescents and young adults (aged 16–25 years) who used cannabis regularly were recruited from the community and a local high school between July 2015 and December 2016. Participants were randomly assigned to 4 weeks of cannabis abstinence, verified by decreasing 11-nor-9-carboxy- $\Delta^9$ -tetrahydrocannabinol urine concentration (MJ-Abst; n = 62), or a monitoring control condition with no abstinence requirement (MJ-Mon; n = 26). Attention and memory were assessed at baseline and weekly for 4 weeks with the Cambridge Neuropsychological Test Automated Battery.

**Results:** Among MJ-Abst participants, 55 (88.7%) met a priori criteria for biochemically confirmed 30-day continuous abstinence. There was an effect of abstinence on verbal memory ( $P = .002$ ) that was consistent across 4 weeks of abstinence, with no time-by-abstinence interaction, and was driven by improved verbal learning in the first week of abstinence. MJ-Abst participants had better memory overall and at weeks 1, 2, 3 than MJ-Mon participants, and only MJ-Abst participants improved in memory from baseline to week 1. There was no effect of abstinence on attention: both groups improved similarly, consistent with a practice effect.

**Conclusions:** This study suggests that cannabis abstinence is associated with improvements in verbal learning that appear to occur largely in the first week following last use. Future studies are needed to determine whether the improvement in cognition with abstinence is associated with improvement in academic and other functional outcomes.

**Trial Registration:** ClinicalTrials.gov identifier: NCT03276221

*J Clin Psychiatry* 2018;79(6):17m11977

<https://doi.org/10.4088/JCP.17m11977>

© Copyright 2018 Physicians Postgraduate Press, Inc.



## Shots

PUBLIC HEALTH

# When Adolescents Give Up Pot, Their Cognition Quickly Improves

October 30, 2018 · 1:01 PM ET

RACHEL D. COHEN



Even a week without marijuana use improves young people's ability to learn and remember.

BURGER/Canopy/Getty Images

Marijuana, it seems, is not a performance-enhancing drug. That is, at least, not among young people, and not when the activity is learning.

A study published Tuesday in the *Journal of Clinical Psychiatry* finds that when adolescents stop using marijuana — even for just one week — their verbal learning and

**memory improve.** The study contributes to growing evidence that marijuana use in adolescents is associated with reduced neurocognitive functioning.

More than 14 percent of students in middle school and high school reported using marijuana within the past month, finds a National Institutes of Health survey conducted in 2017. And marijuana use has increased among high-schoolers over the past 10 years, according to the U.S. Department of Health & Human Services.

At the same time, the percentage of teens who believe that regular marijuana use poses a great risk to their health has dropped sharply since the mid-2000s. And legalization of marijuana may play a part in shaping how young people think about the drug. One study noted that after 2012, when marijuana was legalized in Washington state, the number of eighth-graders there that believed marijuana posed risks to their health dropped by 14 percent.

---

Article continues below

---

## Sign Up For The Health Newsletter

Get the latest stories on the science of healthy living.

SUBSCRIBE

By subscribing, you agree to NPR's terms of use and privacy policy. NPR may share your name and email address with your NPR station. See Details. This site is protected by reCAPTCHA and the Google Privacy Policy and Terms of Service apply.

Researchers are particularly concerned with marijuana use among the young because THC, the active ingredient in marijuana, most sharply affects the parts of the brain that develop during adolescence.

"The adolescent brain is undergoing significant neurodevelopment well into the 20s, and the regions that are last to develop are those regions that are most populated by cannabis receptors and are also very critical to cognitive functioning," says Randi Schuster. Schuster is the director of neuropsychology at Massachusetts General Hospital's Center for Addiction Medicine and the study's lead author.

Schuster and the team of researchers set out to determine if cognitive functions that are potentially harmed by marijuana use in adolescents — particularly attention and memory — improve when they abstain from marijuana.

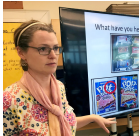
They recruited **88 pot-using teens and young adults, ages 16 to 25,** and got some of them to agree to stop smoking (or otherwise consuming) marijuana for the month.

Schuster says the researchers wanted to recruit a range of participants, not just heavy users or those in a treatment program, for example. Some of the young people smoked once per week; some smoked nearly daily.



#### SHOTS - HEALTH NEWS

Ticket To Ride: Pot Sellers Put Seniors On The Canna-Bus



#### SHOTS - HEALTH NEWS

With The Rise Of Legal Weed, Drug Education Moves From 'Don't' to 'Delay'

The researchers randomly assigned the volunteers into an abstaining group and a nonabstaining group. They delivered the bad news to those chosen to be abstainers at the end of their first visit, and Schuster says, they took it surprisingly well.

"People were generally fine," she says. "We kind of went through what the next month would look like and helped them come up with strategies for staying abstinent."

One motivation for the non-tokers to stick with the program? They received increasing amounts of money each week of the month-long study.

The researchers urine-tested both groups on a weekly basis to make sure that the THC levels for the abstinent group were going down, and that the levels for the control group were staying consistent as they continued using.

Also at each visit, the participants completed a variety of tasks testing their attention and memory through the Cambridge Neuropsychological Test Automated Battery, a validated cognitive assessment tool.

The researchers found that after four weeks, there was no noticeable difference in attention scores between the marijuana users and the nonusers. But, the memory scores of the nonusers improved, whereas the users' memories mostly stayed the same.

The verbal memory test challenged participants to learn and recall new words, which "lets us look both at their ability to learn information the first time the words were presented, as well as the number of words that they're able to retrieve from long-term memory storage after a delay," Schuster says.

Verbal memory is particularly relevant for adolescents and young adults when they're in the classroom, Schuster says.



"For an adolescent sitting in their history class learning new facts for the first time, we're suspecting that active cannabis users might have a difficult time putting that new information into their long-term memory," Schuster says.

While this study didn't prove that abstaining from cannabis improves adolescents' attention, other studies have found that marijuana users fare worse in attention tests than nonusers. Schuster hypothesizes it might take more than four weeks of abstinence for attention levels to improve.

Interestingly, most of the memory improvement for the abstinent group happened during the first week of the study, which leaves the researchers feeling hopeful.

"We were pleasantly surprised to see that **at least some of the deficits that we think may be caused by cannabis appear to be reversible**, and at least some of them are quickly reversible, which is good news," Schuster says.

One weakness of this study is its lack of a non-marijuana-using control group, says Krista Lisdahl, an associate professor of psychology at the University of Wisconsin Milwaukee who was not involved with the study but also researches the neuroscience of addiction. Because of this, it's difficult to conclude whether the improvements in memory brought the participants back to their baseline levels prior to using marijuana.

Also, because the **study lasted only four weeks**, it's impossible to draw conclusions about the long-term effects of marijuana usage for young people, such as how marijuana directly affects academic performance, sleep patterns or mood.

Lisdahl says that longitudinal studies such as the NIH's Adolescent Brain Cognitive Development Study could provide more information about what marijuana does to the adolescent brain. It might also reveal what happens if adolescents stop using marijuana and if their brain functioning can completely recover.

Lisdahl is helping with the NIH study, which has, to date, enrolled more than 11,000 children ages 9 and 10 and will follow them into young adulthood. It's the largest long-term research study on child brain development in the U.S., and it assesses how everything from screen time to concussions to drugs affect adolescents' brains.

In the meantime, Lisdahl says the findings from the new study — that abstinence from marijuana is associated with improvements in adolescents' learning and memory — sends a positive message.

"I remain optimistic that we can show recovery of function with sustained abstinence," she says.



ARTICLES | VOLUME 391, ISSUE 10115, P31-40, JANUARY 06, 2018

## Percutaneous coronary intervention in stable angina (ORBITA): a double-blind, randomised controlled trial

Rasha Al-Lamee, MRCP • David Thompson, MRCPI • Hakim-Moulay Dehbi, PhD • Sayan Sen, MRCP • Kare Tang, FRCP •

John Davies, MRCP • et al. [Show all authors](#) • [Show footnotes](#)Published: November 02, 2017 • DOI: [https://doi.org/10.1016/S0140-6736\(17\)32714-9](https://doi.org/10.1016/S0140-6736(17)32714-9) •

### Summary

#### Background

Symptomatic relief is the primary goal of percutaneous coronary intervention (PCI) in stable angina and is commonly observed clinically. However, there is no evidence from blinded, placebo-controlled randomised trials to show its efficacy.

#### Methods

ORBITA is a blinded, multicentre randomised trial of PCI versus a placebo procedure for angina relief that was done at five study sites in the UK. We enrolled patients with severe ( $\geq 70\%$ ) single-vessel stenoses. After enrolment, patients received 6 weeks of medication optimisation. Patients then had pre-randomisation assessments with cardiopulmonary exercise testing, symptom questionnaires, and dobutamine stress echocardiography. Patients were randomised 1:1 to undergo PCI or a placebo procedure by use of an automated online randomisation tool. After 6 weeks of follow-up, the assessments done before randomisation were repeated at the final assessment. The primary endpoint was difference in exercise time increment between groups. All analyses were based on the intention-to-treat principle and the study population contained all participants who underwent randomisation. This study is registered with [ClinicalTrials.gov](https://clinicaltrials.gov), number [NCT02062593](https://clinicaltrials.gov/ct2/show/study/NCT02062593).

ORBITA enrolled 230 patients with ischaemic symptoms. After the medication optimisation phase and between Jan 6, 2014, and Aug 11, 2017, 200 patients underwent randomisation, with 105 patients assigned PCI and 95 assigned the placebo procedure. Lesions had mean area stenosis of 84·4% (SD 10·2), fractional flow reserve of 0·69 (0·16), and instantaneous wave-free ratio of 0·76 (0·22). There was no significant difference in the primary endpoint of exercise time increment between groups (PCI minus placebo 16·6 s, 95% CI -8·9 to 42·0,  $p=0·200$ ). There were no deaths. Serious adverse events included four pressure-wire related complications in the placebo group, which required PCI, and five major bleeding events, including two in the PCI group and three in the placebo group.

## Interpretation

In patients with medically treated angina and severe coronary stenosis, PCI did not increase exercise time by more than the effect of a placebo procedure. The efficacy of invasive procedures can be assessed with a placebo control, as is standard for pharmacotherapy.

## Funding

NIHR Imperial Biomedical Research Centre, Foundation for Circulatory Health, Imperial College Healthcare Charity, Philips Volcano, NIHR Barts Biomedical Research Centre.

All Content Advanced Search  
☒ Cell Metabolism ☐ All Journals

Explore Online Now Current Issue Archive Journal Information For Authors

&lt; Previous Article

Volume 25, Issue 6, p1243–1253.e5, 6 June 2017

Next Article &gt;

To use the Enhanced view of this article, please enable JavaScript on in your browser and refresh the page.

## CLINICAL AND TRANSLATIONAL REPORT

## Bread Affects Clinical Parameters and Induces Gut Microbiome-Associated Personal Glycemic Responses

Tal Korem<sup>7</sup>, David Zeevi<sup>7</sup>, Niv Zmora, Omer Weissbrod, Noam Bar, Maya Lotan-Pompan, Tali Avnit-Sagi, Noa Kosower, Gal Malka, Michal Rein, Jotham Suez, Ben Z. Goldberg, Adina Weinberger, Avraham A. Levy<sup>7</sup>, Eran Elinav<sup>7</sup>, Eran Segal<sup>8</sup><sup>7</sup> These authors contributed equally<sup>8</sup> Lead contact

Bread Affects Clinical Parameters and Induces Gut Microbiome-Associated Personal Glycemic Responses

DOI: <http://dx.doi.org/10.1016/j.cmet.2017.05.002> |  CrossMark Article Info

Summary Full Text Images References Supp. Info. Comments

## Highlights

- Crossover trial shows no differential clinical effect of white versus sourdough bread
- The microbiome composition was generally resilient to dietary intervention of bread
- The glycemic response to the two types of bread varies greatly across people
- Microbiome-based classifier accurately predicts glycemic-response-inducing bread type

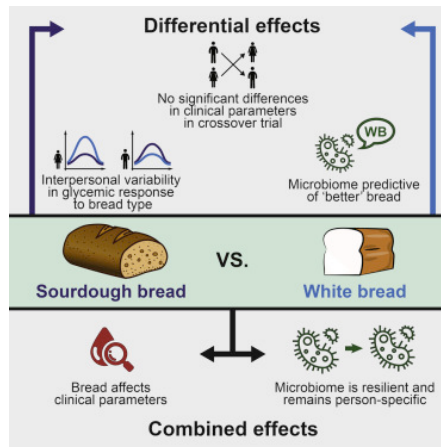
## Summary

Bread is consumed daily by billions of people, yet evidence regarding its clinical effects is contradicting. Here, we performed a randomized crossover trial of two 1-week-long dietary interventions comprising consumption of either traditionally made sourdough-leavened whole-grain bread or industrially made white bread. We found no significant differential effects of bread type on multiple clinical parameters. The gut microbiota composition remained person specific throughout this trial and was generally resilient to the intervention. We demonstrate statistically significant interpersonal variability in the glycemic response to different bread types, suggesting that the lack of phenotypic difference between the bread types stems from a person-specific effect. We further show that the type of bread that induces the lower glycemic response in each person can be predicted based solely on microbiome data prior to the intervention. Together, we present marked personalization in both bread metabolism and the gut microbiome, suggesting that understanding dietary effects requires integration of person-specific factors.

## Keywords:

glycemic responses, gut microbiome, bread, personalization, nutrition, prediction

## Graphical Abstract

PDF (1 MB)  
Extended PDF (1 MB)  
Download Images (ppt)  
Email ArticleAdd to My Reading List  
Export Citation  
Create Citation Alert  
Cited by in Scopus (0)  
Request Permissions  
Order Reprints (100 minimum order)Access this article on  
ScienceDirectC  
Career  
Cel

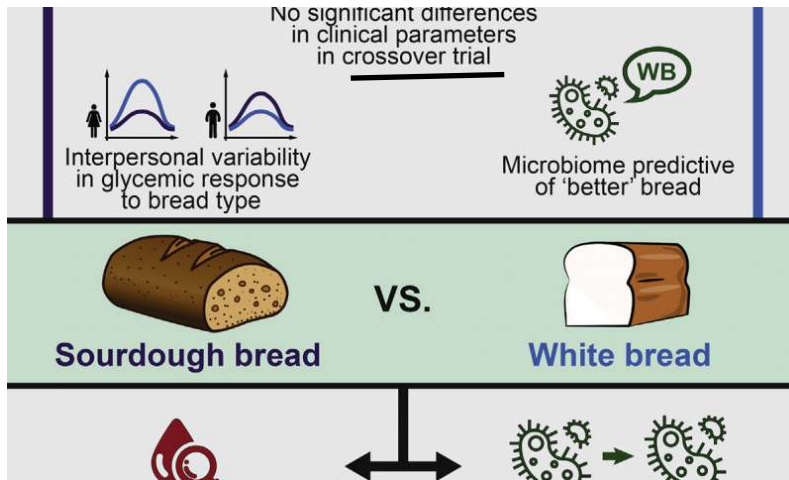
Apj





## Is white or whole wheat bread 'healthier?' Depends on the person

June 6, 2017



This visual abstract shows the findings of Korem et al. who performed a crossover trial of industrial white or artisanal sourdough bread consumption and found no significant difference in clinical effects, with the gut microbiome composition remaining generally stable. They showed the glycemic response to bread type to be person specific and microbiome associated, highlighting the importance of nutrition personalization. Credit: Korem et al./Cell Metabolism 2017

Despite many studies looking at which bread is the healthiest, it is still not clear what effect bread and differences among bread types have on clinically relevant parameters and on the microbiome. In the journal *Cell Metabolism* on June 6, Weizmann Institute researchers report the results of a comprehensive, randomized trial in 20 healthy subjects comparing differences in how processed white bread and artisanal whole wheat sourdough affect the body.

Surprisingly, the investigators found the bread itself didn't greatly affect the participants and that different people reacted differently to the bread. The research team then devised an algorithm to help predict how individuals may respond to the bread in their diets.

All of the participants in the study normally consumed about 10% of their calories from bread. Half were assigned to consume an increased amount of processed, packaged white bread for a week—around 25% of their calories—and half to consume an increased amount of whole wheat sourdough, which was baked especially for the study and delivered fresh to the participants. After a 2-week period without bread, the diets for the two groups were reversed.

Before the study and throughout the time it was ongoing, many health effects were monitored. These included wakeup glucose levels; levels of the essential minerals calcium, iron, and magnesium; fat and cholesterol levels; kidney and liver enzymes; and several markers for inflammation and tissue damage. The investigators also measured the makeup of the participants' microbiomes before, during, and after the study.

"The initial finding, and this was very much contrary to our expectation, was that there were no clinically significant differences between the effects of these two types of bread on any of the

Featured

Last comments

Popular

Vaccinating against psoriasis, allergies and Alzheimer's a possibility, research shows Oct 23, 2017 0

Novel technique explains herbicide's link to Parkinson's disease Oct 23, 2017 0

Activation of immune T cells leads to behavioral changes Oct 23, 2017 0

Neuroscientists build case for new theory of memory formation Oct 23, 2017 1

Exploring how herpes simplex virus changes when passed between family members Oct 22, 2017 0

more »

Medical Xpress on facebook

top

Help

Science X Account

Feature Stories

Android app

Connect

Home

FAQ

Sponsored Account

Latest news

iOS app

Search

About

Newsletter

Week's top

Amazon Kindle

Mobile version

Contact

RSS feeds

Archive

## Is white or whole wheat bread 'healthier?' Depends on the person

they point toward a new paradigm: **different people react differently, even to the same foods,**" says Eran Elinav (@EranElinav), a researcher in the Department of Immunology at the Weizmann Institute and another of the study's senior authors. "To date, the nutritional values assigned to food have been based on minimal science, and one-size-fits-all diets have failed miserably."

He adds: "These findings could lead to a more rational approach for telling people which foods are a better fit for them, based on their microbiomes."

Avraham Levy, a professor in the Department of Plant and Environmental Sciences and another coauthor, adds a caveat to the study: "These experiments looked at everyone eating the same amounts of carbohydrates from both bread types, which means that they ate more whole wheat bread because it contains less available carbohydrates. Moreover, we know that because of its high fiber content, people generally eat less whole wheat [bread](#). We didn't take into consideration how much you would eat based on how full you felt. So the story must go on."

**Explore further:** [Could white bread be making you fat?](#)

**More information:** [Cell Metabolism](#), Korem et al: "Bread affects clinical parameters and induces gut microbiome-associated personal glycemic responses." [http://www.cell.com/cell-metabolism/fulltext/S1550-4131\(17\)30288-7](http://www.cell.com/cell-metabolism/fulltext/S1550-4131(17)30288-7) , DOI: [10.1016/j.cmet.2017.05.002](https://doi.org/10.1016/j.cmet.2017.05.002)

**Journal reference:** [Cell Metabolism](#)

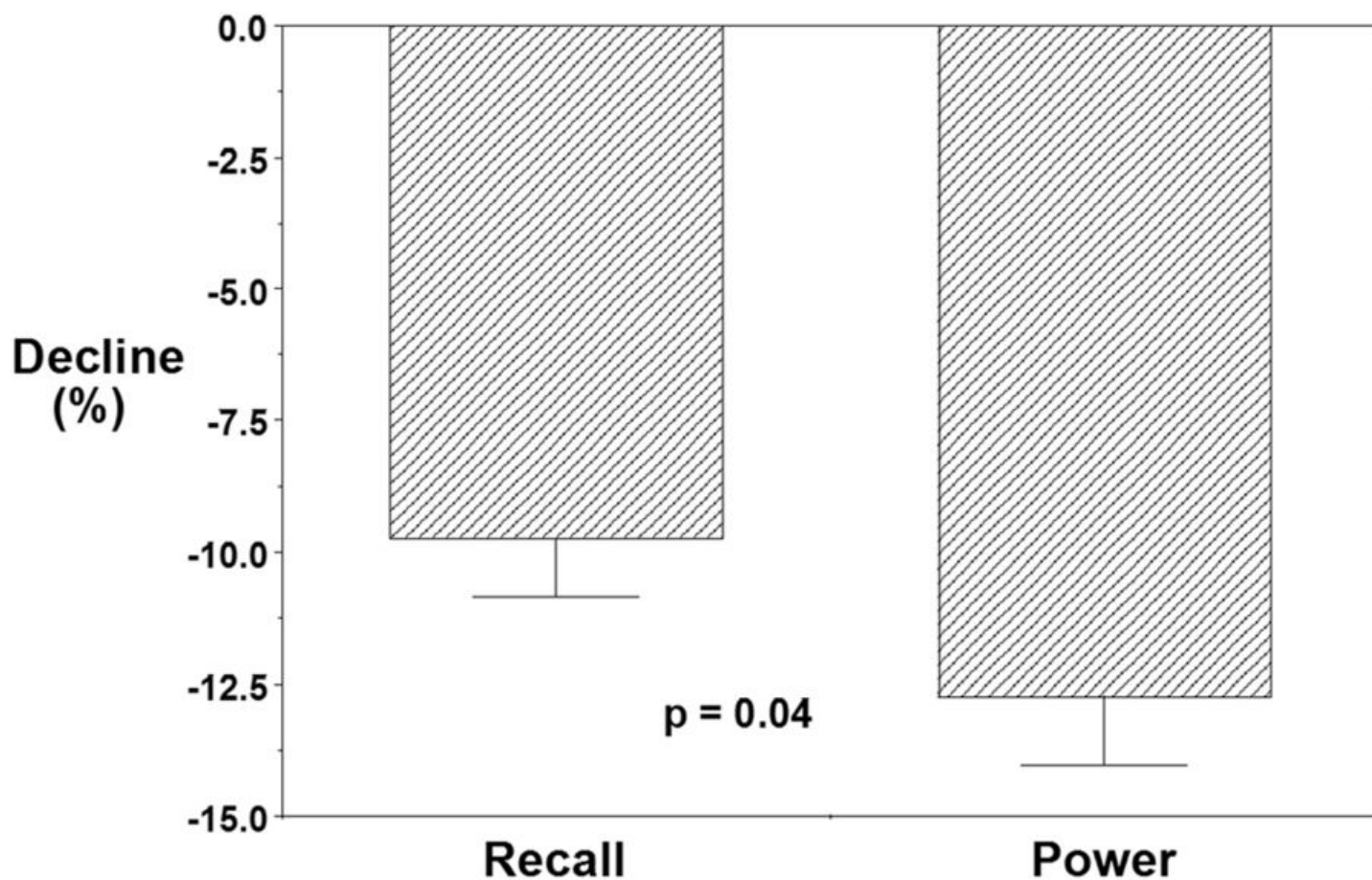
545 shares

**Provided by:** [Cell Press](#)

[feedback to editors](#)

## Figure 1

From: A trade-off between cognitive and physical performance, with relative preservation of brain function



Bar graph showing the relative decrease in cognitive decline and power output. Power output decreased significantly more than cognitive function.

Scientific Reports

[Browse Articles](#)  
[Contact](#)

[Browse Subjects](#)  
[Editor's choice](#)

[About the journal](#)

[Publish](#)

[Journal policies](#)

Journals A-Z

[Nature](#)  
[Scientific Reports](#)

[Nature Communications](#)  
[View all »](#)

[Nature Protocols](#)

[Review journals](#)

All Subjects

[Biological Sciences](#)  
[Physical Sciences](#)

[Earth & Environmental Sciences](#)  
[Scientific Community & Society](#)

[Health Sciences](#)  
[View all »](#)

# SCIENTIFIC REPORTS

OPEN

## A trade-off between cognitive and physical performance, with relative preservation of brain function

Daniel Longman<sup>1</sup>, Jay T. Stock<sup>1,2</sup> & Jonathan C. K. Wells<sup>3</sup>

Received: 21 July 2016

Accepted: 19 September 2017

Published online: 20 October 2017

Debate surrounds the issue of how the large, metabolically expensive brains of *Homo sapiens* can be energetically afforded. At the evolutionary level, decreased investment in muscularity, adiposity and the digestive tract allow for a larger brain. Developmentally, high neo-natal adiposity and preferential distribution of resources to the brain provide an energetic buffer during times of environmental stress. Through an experimental design, we investigated the hypothesis of a trade-off involving brain and muscle at the acute level in humans. Mental performance was measured by a free-recall test, and physical performance by power output on an indoor rowing ergometer. Sixty-two male student rowers performed the two tests in isolation, and then again simultaneously. Paired samples *t*-tests revealed that both power output and mental performance reduced when tested together compared to in isolation ( $t(61) = 9.699$ ,  $p < 0.001$  and  $t(61) = 8.975$ ,  $p < 0.001$ ). Furthermore, the decrease in physical performance was greater than the decrease in mental performance ( $t(61) = -2.069$ ,  $p = 0.043$ ). This is the first investigation to demonstrate an acute level trade-off between these two functions, and provides support for the selfish brain hypothesis due to the relative preservation of cognitive function over physical power output. The underlying mechanism is unclear, and requires further work.

**Evolutionary and developmental implications of enhanced encephalization.** The development of an enlarged and elaborated brain is considered a defining characteristic of human evolution<sup>1</sup>. The evolution of the *Homo* clade has been accompanied by significant encephalization<sup>2,3</sup>. This facilitated the development of more complex social strategies<sup>4,5</sup>, more effective food acquisition<sup>6</sup> and the ability to solve ecological problems through innovative means<sup>7</sup>. Each of these characteristics may have increased survival and reproductive success, giving a greater life expectancy at the age of first reproduction<sup>8</sup>.

While the benefits of encephalization are numerous, the brain imposes significant metabolic costs on both the individual<sup>9–11</sup>. High levels of energetic expenditure are necessitated by the brain's responsibility for regulating the body's energy supply and controlling the function of many peripheral organs<sup>12</sup>. These functions require intense neuronal activity, giving the brain the highest metabolic demand relative to size of all organs<sup>13</sup>.

The question of how larger brains can be metabolically afforded has remained a prominent problem in human evolution<sup>11,14–17</sup>. Life history theory states that as energy availability is finite, an organism has a limited energy budget. Energy allocated to one function cannot be used for another. Energy savings in other organs or tissues could allow for energetic diversion to the brain, without the need to increase overall metabolic expenditure<sup>11,18</sup>. Such a trade-off has been proposed with both digestive tract development<sup>17</sup> and adiposity<sup>19</sup>.

**Meeting the brain's metabolic requirements.** The immediate metabolic costs of the brain depend on its activation state. While the metabolic rate is low during sleep<sup>20</sup> increased energy consumption has been observed in response to a mental task<sup>21</sup>, and following somatosensory, olfactory, visual and auditory stimulation<sup>22–27</sup>. The adult brain almost exclusively derives its energy from the metabolism of glucose<sup>28</sup>. This, coupled with its high energetic demand, ensure that the brain metabolises the most glucose of any organ<sup>29,30</sup>. The brain, however, is unable to store significant amounts of energy and hence buffer its high yet variable metabolic demand<sup>31</sup>. As such, the body is required to supply glucose to the brain quickly and effectively. The 'Selfish Brain Hypothesis'<sup>12</sup> posits that the brain prioritises its own glucose needs over those of the peripheral organs, such as skeletal muscle.

<sup>1</sup>Department of Archaeology and Anthropology, University of Cambridge, Cambridge, CB2 3QG, UK. <sup>2</sup>Department of Anthropology, University of Western Ontario, Ontario, Canada. <sup>3</sup>Childhood Nutrition Research Centre, UCL Institute of Child Health, London, WC1N 1EH, UK. Correspondence and requests for materials should be addressed to D.L. (email: [dl329@cam.ac.uk](mailto:dl329@cam.ac.uk))



Experimental protocol summary	
Protocol	Description
A	Physical task
B	Mental task
C	Physical and mental task

**Table 3.** Experimental protocols.

was recorded. Protocol C consisted of the same 3 minute row as A, but while simultaneously performing the mental task of Protocol B. Both average power output and number of words correctly recalled were recorded.

The rowing ergometer was used because it is an energetically demanding activity, and has been used in previous studies investigating extreme physical stress<sup>77,78</sup>. The mental task involved free recall. A large printed screen showing 75 words was clearly displayed in front of the participants' chair (Protocol B), or in front of the rowing ergometer (Protocol C), for a duration of 3 minutes. The participants were required to recall and write as many words as possible in any order from memory within 5 minutes (5 minutes immediately following the row in Protocol C)<sup>79</sup>. The words were selected from the Toronto Noun Pool<sup>80</sup>. Two 75-word lists were randomly created from the 150 words used by Kahana & Howard<sup>81</sup> and were counterbalanced across participants. Half of the participants were given List 1 for Protocol B and List 2 for Protocol C, with the other half being given List 2 for Protocol B and List 1 for Protocol C. This method ensured that each word was seen an equal number of times across participants, and each participant saw each word only once. Such counterbalancing ensured that any artefacts<sup>82</sup> were controlled for to reduce the likelihood of such artefacts<sup>83</sup>.

The Protocols were completed at 1 week intervals. All participants refrained from extra exercise the day before, and the day of, each Protocol. The same machine was used for Protocols A and B, with the drag factor being consistent. The order in which the participants completed the three protocols was also counterbalanced, in order to control for any effects such as the development of memorising strategies.

## References

- Foley, R. & Lee, P. Ecology and energetics of encephalization in hominid evolution. *Philos Trans R Soc Lond B Biol Sci.* **334**(1270), 223–32 (1991).
- Hawks, J., Hunley, K., Lee, S. H. & Wolpoff, M. Population bottlenecks and Pleistocene human evolution. *Mol Biol Evol.* **17**(1), 2–22 (2000).
- Lee, S. & Wolpoff, M. The pattern of evolution in Pleistocene human brain size. *Paleobiology.* **29**(2), 186–96 (2003).
- Byrne, R. & Corp, N. Neocortex size predicts deception rate in primates. *Proc Biol Sci.* **271**(1549), 1693–9 (2004).
- Parker, S. & McKinney, M. Origins of intelligence: the evolution of cognitive development in monkeys, apes, and humans. (Johns Hopkins University Press, 1999).
- Gibson, K. Cognition, brain size and the extraction of embedded food resources. *Primate Ontog Cogn Soc Behav.* **3**, 92–10 (1986).
- Reader, S. & Laland, K. Social intelligence, innovation, and enhanced brain size in primates. *Proc Natl Acad Sci USA* **99**(7), 4436–41 (2002).
- Barrickman, N., Bastian, M., Isler, K. & van Schaik, C. Life history costs and benefits of encephalization: a comparative test using data from long-term studies of primates in the wild. *J Hum Evol.* **54**(5), 568–90 (2008).
- Mink, J., Blumenshine, R. & Adams, D. Ratio of central nervous system to body metabolism in vertebrates: its constancy and functional basis. *Am J Physiol Integr Comp Physiol.* **241**(3), R203–12 (1981).
- Bullmore, E. & Sporns, O. The economy of brain network organization. *Nat Rev Neurosci.* **13**(5), 336–49 (2012).
- Isler, K. & van Schaik, C. P. Metabolic costs of brain size evolution. *Biol Lett.* **2**(4), 557–60 (2006).
- Peters, A. *et al.* The selfish brain: competition for energy resources. *Neurosci Biobehav Rev.* **28**, 143–80 (2004).
- Attwell, D. & Laughlin, S. An energy budget for signalling in the grey matter of the brain. *J Cereb Blood Flow Metab.* **21**(10), 1133–45 (2001).
- Aiello, L. & Dunbar, R. Neocortex size, group size, and the evolution of language. *Curr Anthropol.* **34**(2), 183–93 (1993).
- Byrne, R. The technical intelligence hypothesis: an additional evolutionary stimulus to intelligence? In *Machiavellian intelligence II: extension and evaluations* (eds Whiten, A., Byrne, R.) 289–311 (Cambridge University Press, 1997).
- Isler, K. & van Schaik, C. Costs of encephalization: the energy trade-off hypothesis tested on birds. *J Hum Evol.* **51**(3), 228–43 (2011).
- Aiello, L. & Wheeler, P. The expensive-tissue hypothesis the brain and the digestive evolution. *Curr Anthropol.* **36**(2), 199–221 (1995).
- McNab, B. & Eisenberg, J. F. Brain size and its relation to the rate of metabolism in mammals. *Am Nat.* **133**(2), 157–67 (1989).
- Navarrete, A., van Schaik, C. P. & Isler, K. Energetics and the evolution of human brain size. *Nature.* **480**(7375), 91–3 (2011).
- Madsen, P. & Vorstrup, S. Cerebral blood flow and metabolism during sleep. *Cerebrovasc brain met.* **3**(4), 281–96 (1990).
- Madsen, P. *et al.* Persistent resetting of the cerebral oxygen/glucose uptake ratio by brain activation: evidence obtained with the Key-Schmidt technique. *J Cereb blood flow Metab.* **15**, 485–91 (1995).
- Fox, P. & Raichle, M. Focal physiological uncoupling of cerebral blood flow and oxidative metabolism during somatosensory stimulation in human subjects. *Proc Natl Acad Sci USA* **83**, 1140–4 (1986).
- Fox, P., Raichle, M., Mintun, M. & Dence, C. Nonoxidative glucose consumption during focal physiological neural activity. *Science.* **241**(4864), 462–4 (1988).
- Sharp, F., Kauer, J. & Shepherd, G. Local sites of activity-related glucose metabolism in rat olfactory bulb during olfactory stimulation. *Brain Res.* **98**(3), 596–600 (1975).
- Kennedy, C. *et al.* Metabolic mapping of the primary visual system of the monkey by means of the autoradiographic [<sup>14</sup>C] deoxyglucose technique. *Proc Natl Acad Sci USA* **73**(11), 4230–4 (1976).
- Ginsberg, M. D., Dietrich, W. D. & Busto, R. Coupled forebrain increases of local cerebral glucose utilization and blood flow during physiologic stimulation of a somatosensory pathway in the rat: demonstration by double-label autoradiography. *Neurology.* **37**(1), 11–9 (1987).
- Roland, P., Eriksson, L., Stone-Elander, S. & Widen, L. Does mental activity change the oxidative metabolism of the brain? *J Neurosci.* **7**, 2373–89 (1987).
- Bélanger, M., Allaman, I. & Magistretti, P. Brain energy metabolism: focus on astrocyte-neuron metabolic cooperation. *Cell Metab.* **14**(6), 724–38 (2011).

# The Crossover Study

potential outcomes realized

The study in Chapter 9 comparing four formulas for their effects on the gains in weight of infants is an example of a crossover study, an experiment in which subjects are administered first one treatment, then “crossed over” to receive a second, and perhaps subsequently crossed over to receive a third or even a fourth. The intuitive appeal of having each subject serve as his or her own control has made the crossover study one of the most popular experimental strategies since the infancy of formal experimental design [see Federer (1955, p. 438) for references to some early applications]. Frequent misapplications of the design in clinical experiments, and frequent misanalyses of the data, motivated the Biometric and Epidemiological Methodology Advisory Committee to the U.S. Food and Drug Administration to recommend in June of 1977 that, in effect, the crossover design be avoided in comparative clinical studies except in the rarest instances.

This introductory section to a chapter mainly devoted to methods for analyzing the data from crossover studies is the appropriate place to discuss when the design may be in order and when not (see also Brown, 1980; Fisher and Wallenstein, 1981; Grizzle, 1965; and Hills and Armitage, 1979). The design obviously is not in order when the condition or disease being treated is acute (e.g., postoperative pain, depression in reaction to the loss of a loved one, and the common cold), because the chances are high that there will be nothing left to treat when the patient is scheduled to receive the second or later treatments. The crossover design should be reserved for studies of treatments for chronic conditions (e.g., angina pectoris, so-called “endogenous” depression, and essential hypertension).

The design should also be reserved for those treatments whose effects can be measured after only a “short” period of administration. The word *short* was included within quotation marks because it necessarily means different durations for different conditions and to different people. For a condition that a patient has suffered from for many years, six months of

# Repeated measures designs

## Cross-over Designs

- Subjects receive every treatment
- Most common is ``two-period, two-treatment"
  - Subjects are randomly assigned to receive either
    - A in period 1, B in period 2 or
    - B in period 1, A in period 2

# Repeated measures designs

## Cross-over Designs

- Important assumption: No carry-over effects
  - effect of treatment received in each period is not affected by treatment received in previous periods.
- To minimize possibility of carry-over effects
  - ‘wash-out’ time between the periods in which treatments are received.



# Cross-over designs: Example

- Treatments: Impermeable (IP) / Semi-Permeable (SP)
- Outcomes: Skin temperature, heat storage, oxygen consumption
- Protocol:
  - 6 men studied under both types of clothing.
  - 3 men randomized to order (IP, SP), 3 men to (SP, IP)

Rissanen and Rintamaki (1997) Ergonomics p. 141-150.

## CROSSOVER DESIGNS

So far, we have considered the single-group repeated measures design where each subject receives each of  $p$  treatments at  $p$  different occasions.

Next we consider a variant of the single-group repeated measures design known as the crossover design.

In the simplest version of the cross-over design, two treatments, say A and B, are to be compared. Subjects are randomly assigned to the two treatment orders:  $A \rightarrow B$  and  $B \rightarrow A$ .

**Example:** Placebo-controlled study of the effect of erythropoietin on plasma histamine levels and pruritus scores of 10 dialysis patients.

Treatment schedule was 5 weeks of placebo and 5 weeks of erythropoietin in random order.

Designs in which subjects are randomly assigned to either  $P \rightarrow T$  (placebo, treatment) or  $T \rightarrow P$  are called two-period crossover designs.

If we assume that there is no carryover<sup>2</sup> of treatment effects from period 1 to period 2, we can write the basic model as

$$Y_{ij} = \beta_0 + \beta_1 \text{time}_{ij} + \beta_2 \text{trt}_{ij} + e_{ij}$$

where  $Y_{ij}$  is the response of subject  $i$  at time  $j$ , and  $\text{time}_{ij}$  and  $\text{trt}_{ij}$  are the values of the time and treatment variable associated with  $Y_{ij}$ .

If there is a carryover of the effect of treatment (e.g. erythropoietin) from period 1 to period 2, we need to define a new indicator variable:

$$CO_{ij} = \begin{cases} 1, & \text{if } T \text{ given in the previous period;} \\ 0, & \text{otherwise.} \end{cases}$$

This indicator variable will equal 1 only in the second period for the group assigned to T→P.

---

<sup>2</sup>**Carryover** is the persistence of a treatment effect applied in one period in a subsequent period of treatment.

**Example: Placebo-controlled study of the effect of erythropoietin on plasma histamine levels.**

id	time	trt	co	y
1	1	1	0	24
1	2	2	0	5
2	1	1	0	23
2	2	2	0	8
3	1	1	0	19
3	2	2	0	3
4	1	1	0	26
4	2	2	0	8
5	1	1	0	16
5	2	2	0	3
6	1	2	0	2
6	2	1	1	18
7	1	2	0	8
7	2	1	1	29
8	1	2	0	5
8	2	1	1	26
9	1	2	0	6
9	2	1	1	28
10	1	2	0	4
10	2	1	1	19

# Longitudinal Designs

Crossover design example

no carryover (CO=0)

$$Y_{ij} = \beta_0 + \beta_1 \text{time}_{ij} + \beta_2 \text{trt}_{ij} + e_{ij}$$

data slide 137 Harvard ALA see linked material

```
> cross = read.table(file="D:\\drr12\\stat222\\week6\\crosspl37.dat", header = T)
```

```
> head(cross)
```

10 subjects 2 trts

```
  id time trt co y
1  1     1   1  0 24
2  1     2   2  0  5
3  2     1   1  0 23
4  2     2   2  0  8
5  3     1   1  0 19
6  3     2   2  0  3
```

```
> attach(cross)
```

repeated measures error term

```
> craov = aov(y ~ as.factor(time) + as.factor(trt) + Error(as.factor(id)))
```

anova model

```
> summary(craov)
```

could use  
lmer  
instead  
RQ5

Error: as.factor(id)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	9	189		21	

Error: Within

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
as.factor(time)	1	9.8	9.8	2.42	0.158
as.factor(trt)	1	1548.8	1548.8	382.42	4.86e-08 ***
Residuals	8	32.4	4.1		

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
> sqrt(382.42)
```

```
[1] 19.55556 ← t for fixed effects in SAS
```

```
> #matches ALA slide 144, no carryover assumed CO=0
```

cf ecg data from ALA

# data slide 137 Harvard ALA

```
> cross = read.table(file="D:\\drr12\\stat222\\week6\\crossp137.dat", header = T)
> head(cross)
  id time trt co  y
1  1    1   1  0 24
2  1    2   2  0  5
3  2    1   1  0 23
4  2    2   2  0  8
5  3    1   1  0 19
6  3    2   2  0  3
> attach(cross)
> craov = aov(y ~ as.factor(time) + as.factor(trt) + Error(as.factor(id)))
> summary(craov)
```

```
Error: as.factor(id)
      Df Sum Sq Mean Sq F value Pr(>F)
Residuals  9    189      21
```

```
Error: Within
      Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(time)  1    9.8      9.8    2.42    0.158
as.factor(trt)   1 1548.8  1548.8  382.42 4.86e-08 ***
Residuals       8   32.4      4.1
```

---

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> sqrt(382.42)
```

```
[1] 19.55556
```

```
> #matches ALA slide 144, no carryover assumed
```



### Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr >  T
Intercept	5.9000	1.2062	9	4.89	0.0009
time 1	-1.4000	0.9000	8	-1.56	0.1584
time 2	0	.	.	.	.
trt 1	17.6000	0.9000	8	19.56	<0.0001
trt 2	0	.	.	.	.

### Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	Chi-Square	Pr > ChiSq
time	1	8	2.42	0.1198
trt	1	8	382.42	0.0001

The pooled estimate of the treatment effect, combining the responses from period 1 and period 2, is 17.6.

Note: Assuming no carryover effects, the treatment effect is based on within-subject comparisons, and the standard error has decreased from 2.33 to 0.90.

With no carryover effects, the treatment effect is estimated by

$$(\bar{d}_1 + \bar{d}_2) / 2, \text{ which has variance } \sigma_w^2 / n.$$

In contrast, with two independent groups, and  $2n$  subjects in each group, an estimate of the treatment effect has variance  $(\sigma_b^2 + \sigma_w^2) / n$ .

Thus, the crossover design has the potential to substantially increase the precision of the estimate of the treatment effect.

Data from a two-period crossover trial on cerebrovascular deficiency.

Source: Table 3.1 (page 90) of Jones and Kenward (1989).  
With permission of CRC Press.

Reference: Jones, B. and Kenward, M.G. (1989). Design and Analysis of Cross-over Trials.  
London: Chapman and Hall/CRC Press.

#### Description:

The dataset consists of safety data from a crossover trial on the disease cerebrovascular deficiency. The response variable is not a trial endpoint but rather a potential side effect. In this two-period crossover trial, comparing the effects of active drug to placebo, 67 patients were randomly allocated to the two treatment sequences, with 34 patients receiving placebo followed by active treatment, and 33 patients receiving active treatment followed by placebo. The response variable is binary, indicating whether an electrocardiogram (ECG) was abnormal (Y=1) or normal (Y=0). Each patient has a bivariate binary response vector.

#### Variable List:

Subject ID, Sequence (1=Placebo-->Active, 2=Active-->Placebo), Period (0=Period 1, 1=Period 2),  
TRT (0=Placebo, 1=Active Drug), ECG Response (0=Normal, 1=Abnormal).

1	1	0	0	0
1	1	1	1	0
2	1	0	0	0
2	1	1	1	0
3	1	0	0	0
3	1	1	1	0
4	1	0	0	0
4	1	1	1	0
5	1	0	0	0
5	1	1	1	0
6	1	0	0	0
6	1	1	1	0
7	1	0	0	0
7	1	1	1	0
8	1	0	0	0
8	1	1	1	0
9	1	0	0	0
9	1	1	1	0
10	1	0	0	0
10	1	1	1	0
11	1	0	0	0
11	1	1	1	0
12	1	0	0	0
12	1	1	1	0
13	1	0	0	0
13	1	1	1	0
14	1	0	0	0
14	1	1	1	0
15	1	0	0	0
15	1	1	1	0
16	1	0	0	0
16	1	1	1	0
17	1	0	0	0
17	1	1	1	0
18	1	0	0	0
18	1	1	1	0
19	1	0	0	0
19	1	1	1	0
20	1	0	0	0
20	1	1	1	0
21	1	0	0	0

# ANALYSIS OF CROSS-OVER DESIGN

## Duration Effect of Three Formulations of a Drug

Twelve males volunteered to participate in a study to compare the effect of three formulations of a drug product: Formulation 1 was a 5-mg tablet, Formulation 2 was a 100-mg tablet, and Formulation 3 was a sustained-release capsule. There are  $3! = 6$  possible sequences in which the three formulations could be administered to the subjects during the three treatment periods.

Sequence	Time Period		
	1	2	3
1	$F_1$	$F_2$	$F_3$
2	$F_2$	$F_3$	$F_1$
3	$F_3$	$F_1$	$F_2$
4	$F_2$	$F_1$	$F_3$
5	$F_3$	$F_2$	$F_1$
6	$F_1$	$F_3$	$F_2$

The experimenter selected the first 3 of the 6 sequences and randomly assigned 4 subjects to each sequence. The following model describes the study:

$$Y_{ijk} = \mu + \alpha_i + b_{j(i)} + \gamma_k + \tau_{d(i,k)} + \epsilon_{ijk},$$

where  $\alpha_i$ ,  $i = 1, 2, 3$  - sequence effect;  $b_{j(i)}$ ,  $j = 1, 2, 3, 4$  - patient within sequence effect;  $\gamma_k$ ,  $k = 1, 2, 3$ ; - Time period effect  $\tau_{d(i,k)}$ ,  $d = 1, 2, 3$  - Treatment Effect;  $\epsilon_{ijk}$  - experimental error effect

On each treatment day, volunteers were given their assigned formulation and were observed to determine the duration of effect (blood pressure lowering). The experimental data is given here.

Sequence	Patient(Seq)	Time Period			P(S)	Sequence
		1	2	3	$\bar{Y}_{ij.}$	$\bar{Y}_{i..}$
1	1	1.5	2.2	3.4	2.36	2.383
	2	2.0	2.6	3.1	2.56	
	3	1.6	2.7	3.2	2.5	
	4	1.1	2.3	2.9	2.1	
2	1	2.5	3.5	1.9	2.36	2.383
	2	2.8	3.1	1.5	2.56	
	3	2.7	2.9	2.4	2.5	
	4	2.4	2.6	2.3	2.1	
3	1	3.3	1.9	2.7	2.36	2.383
	2	3.1	1.6	2.5	2.56	
	3	3.6	2.3	2.2	2.5	
	4	3.0	2.5	2.0	2.1	
Time	$\bar{Y}_{..k}$	2.46	2.516	2.5083	$\bar{Y}_{...} = 2.497$	
Formulation	$\bar{Y}_d$	1.883	2.46	3.1416	$\bar{Y}_{...} = 2.497$	

The general setting of a crossover design will now be described. Suppose we have  $t$  treatments which are to be compared with respect to their mean responses. In the experiment we have either very heterogeneous EU's or a limited number of EU's and decide that each EU will be observed under all  $t$  treatments. The EU's serve as blocks and thus control the variation in response from EU to EU for a given treatment. An obvious question of concern is whether or not the order in which the EU receives the treatments has an effect on the responses. There are  $t!$  possible sequences in which the  $t$  treatments may be applied. Generally only a subset of the  $t!$  possible sequences will be used in the study. The experimenter decides on  $n$  sequences which are of greatest interest. There will be  $r_i$  EU's randomly assigned to the  $i$ th treatment sequence which will be observed during  $p$  time periods. There is generally a time delay between administering the treatments and when the response is measured on the EU. Furthermore, after the measurements are taken, there will be a further delay before the next treatment is applied in order that the effect of the previously administered treatment not have a *carryover effect* on the EU during the administering of the next treatment. This is called the *washout period*. The following model would be applicable:

$$Y_{ijkdc} = \mu + \alpha_i + b_{j(i)} + \gamma_k + \tau_{d(i,k)} + \lambda_{c(i,k)} + e_{ijk}$$

where  $i = 1, \dots, n$ ;  $j = 1, \dots, r_i$ ;  $k = 1, \dots, p$ ;  $d = 1, \dots, t$  and  $\mu$  is the overall mean response,  $\alpha_i$  is the effect of the  $i$ th sequence,  $b_{j(i)}$  is the random effect for the  $j$ th EU in the  $i$ th sequence,  $\gamma_k$  is the  $k$ th time period effect,  $\tau_{d(i,k)}$  is the direct effect of the treatment applied during period  $k$  in sequence  $i$  and  $\lambda_{c(i,k)}$  is the carryover effect of the treatment applied during period  $k$  in sequence  $i$ . Note that there is randomization of the subjects to the sequences. Furthermore, there are two sizes of EU's. The **EU for Sequence is "Subject"** and the **EU for Treatment is "Time Period"**. The Sequence effect measures some form of the Time Period by Treatment Interaction and may be an indication of a Carryover Effect and/or Correlation in the measurements over Time Periods. When the Sequence effect is highly significant, only the data from the **First Time Period** is used in testing for Treatment effects.

A particular unique characteristic of the Crossover Design is that each Subject receives all  $t$  Treatments. A degree of balance is obtained in the crossover design by having each treatment follow every other treatment the same number of times in the study, each treatment occurs the same number of times in each time period, and each treatment is observed only once on each EU. These characteristics create some particular advantages and disadvantages for the Crossover Design:

**Advantages:**

1. Reduction in the Between EU variation (Subject is serving as a blocking variable)
2. Increases the precision in comparing treatment means
3. Reduction in experimental cost when EU's are expensive and/or difficult to recruit for study and/or difficult/expensive to maintain during study.

**Disadvantages:**

1. May be a carryover effect which will invalidate much of the study
2. Reduced information/coverage of the population of EU's

There is a further complication with the above model besides the potential of the carryover effect. There are  $t$  observations on each EU under the  $t$  different treatments. Thus, we have a multivariate response on each EU, not a single response. Under special conditions, which were discussed in the **Repeated Measures section** of this course, we can validly analyze the data as an univariate experiment. Furthermore, if there was not a carryover effect then we could analyze the experiment as a Latin Square Design with Blocking Variables: Sequence and Time Period. In the previous example, we did not consider the carryover effect in the model. In order to include this term in the model it is necessary to run several models and determine the change in sum of squares for error due to excluding particular terms from the model. In order to accomplish this we must run PROC GLM in order to obtain all the pertinent sums of squares:



# crossdes

## A Package for Design and Randomization in Crossover Studies

by Oliver Sailer

### Introduction

Design of experiments for crossover studies requires dealing with possible order and carryover effects that may bias the results of the analysis. One approach to deal with those effects is to use designs balanced for the respective effects. Almost always there are effects unaccounted for in the statistical model that may or may not be of practical importance. An important way of addressing those effects is randomization.

**crossdes** constructs and randomizes balanced block designs for multiple treatments that are also balanced for carryover effects. Jones and Kenward (1989), Ch. 5 review different crossover designs and cite optimality results. Wakeling and MacFie (1995) promoted the use of balanced designs for sensory studies. Five construction methods described in this paper are implemented here. They include Williams designs (Williams, 1949) and designs based on complete sets of mutually orthogonal latin squares. If one is just interested in getting balanced incomplete block designs that are not balanced for carryover, the package **AlgDesign** may be used (Wheeler, 2004). **crossdes** also contains functions to check given designs for balance.

**crossdes** is available on CRAN. It requires three packages also available on CRAN: **AlgDesign**, **gtools**, located in package bundle **gregmisc**, and **MASS**, available in bundle **VR**.

### Simple Crossover Studies for Multiple Treatments

When comparing the effects of different treatments on experimental units, economical and ethical considerations often limit the number of units to be included in the study. In crossover studies, this restraint is addressed by administering multiple treatments to each subject.

However, the design setup leads to a number of possible problems in such an analysis. In addition to the treatment effects, we have to consider subject effects, order effects and carryover effects. In case we are explicitly interested in estimating those effects, we may fit corresponding mixed effects models. If, however, we are only interested in differences between treatments, we may use balanced designs that average out the nuisance parameters as far as possible.

By far the most common crossover design is the AB/BA design with just two treatments applied in two periods, see e.g. Jones and Kenward (1989) or Senn (1993). Here we restrict our attention to balanced designs for the comparison of three or more treatments and to cases, where each unit receives each treatment at most once. In general these designs are no longer balanced for treatment or carryover effects if some observations are missing. Simulation studies suggest that these designs are fairly robust against small numbers of dropouts, see e.g. Kunert and Sailer (2005). In the next section we explain how to actually get balanced designs in R using **crossdes**.

### Design Construction

There are three important parameters that we need to define before we can construct a design: The number of treatments  $t$  to compare, the number of periods  $p \leq t$ , i.e. the number of treatments each subject gets and the (maximum) number of subjects or experimental units  $n$  available for the study.

To ways to get balanced designs are implemented. The first approach is to specify one of the five construction methods described below. Unless there is no design for the specified parameters  $t$ ,  $p$  and  $n$ , a matrix representing the experimental plan is given. Rows represent subjects and columns represent periods. The design should then be randomized. The function `random.RT` randomizes row and treatment labels.

The function `all.combin` constructs balanced block designs that are balanced for all orders of carryover effects up to  $p - 1$  based on all possible permutations of treatment orders. The user specifies the number of treatments and periods,  $t$  and  $p$ . While this approach works for any  $t \geq 2$  and  $p \leq t$ , the fact that every treatment combination occurs leads to very large numbers of subjects required for the study as long as  $t$  and  $p$  aren't very small, see e.g. Patterson (1952).

As long as  $t$  is a prime power, there is a possibility to drastically reduce the number of subjects required and still retain the same properties as described above. The function `des.MOLS` constructs designs based on complete sets of mutually orthogonal latin squares (MOLS) where  $t \leq 100$  has to be a prime power and  $p \leq t$ , see e.g. Williams (1949). The function `MOLS` gives a complete set of  $t - 1$  MOLS based on Galois Fields that is of use in other applications of combinatorics in the design of experiments as well (Street and Street, 1987). The necessary arguments are  $r$  and  $s$ , where  $r$  is prime and  $s$  is a positive integer such that  $r^s = t$ .

If the number of subjects is still too large or  $t$  is not a prime power, one may consider designs that are

only balanced for first order carryover effects. If each subject gets each treatment once, Williams designs only use  $t$  or  $2t$  subjects (Williams, 1949). `williams` gives carryover balanced latin squares for  $t \geq 2$ .

If the number of treatments is large, it may not be possible for the subjects to receive all treatments. Two construction methods for incomplete designs are implemented in the package. One is `williams.BIB`, which combines balanced incomplete block designs (BIBD) and Williams designs (Patterson, 1951). The user needs to specify a balanced incomplete block design. Such designs may be found using the package `AlgDesign` (Wheeler, 2004). The function `find.BIB` provides a convenient way to use that package to get a BIBD. The last construction function considered here is `balmin.RMD`, a function that constructs the balanced minimal repeated measurements designs of Afsarinejad (1983). These designs are balanced for first order carryover effects but in general, they are not balanced for subject and order effects.

A more convenient way is to use the menu driven function `get.plan` documented below. The user specifies  $t$ ,  $p$  and the maximum number of subjects available. The function checks which of the above functions may work for the given parameters. If there is no design that fits to the parameters, the design parameters may be adapted and a new choice of methods is presented.

For example, we may want to get a design for the comparison of 7 products. Assume that a maximum of 100 test persons are available and they may test all products within one session, i.e.  $t = p = 7$  and  $n = 100$ . We have

```
> design <- get.plan(7,7,100)
```

Possible constructions and minimum numbers of subjects:

	1	2
Method:	<code>williams</code>	<code>des.MOLS</code>
Number:	14	42

Please choose one of the following constructions

```
1:williams
2:des.MOLS
3:Exit
```

Selection:

`get.plan` suggests to use either a Williams design (which requires 14 subjects only) or to use a set of MOLS, which requires 42 subjects. Assume that we are interested in keeping the number of subjects as low as possible. We choose a williams design and type in the corresponding number:

```
> Selection: 1
williams selected. How many 'replicates'
do you wish (1 - 7)?
```

Selection:

We choose not to replicate the design since that would mean additional subjects and type in 1. If we wanted to get closer to the maximum number of subjects available we could have selected a design based on MOLS (for higher order balance) and two replicates of the design. That would have meant 84 subjects.

```
> Selection: 1
1 replicate(s) chosen
Row and treatment labels have been
randomized.
Rows represent subjects, columns
represent periods.
```

As indicated, the design is already randomized. A possible realization of the treatment sequences for the subjects is as follows:

```
> design
      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,] 1 3 4 5 7 6 2
[2,] 3 5 1 6 4 2 7
[3,] 5 6 3 2 1 7 4
[4,] 7 2 4 6 1 5 3
[5,] 3 1 5 4 6 7 2
[6,] 1 4 3 7 5 2 6
[7,] 7 4 2 1 6 3 5
[8,] 5 3 6 1 2 4 7
[9,] 2 7 6 4 5 1 3
[10,] 6 2 5 7 3 4 1
[11,] 6 5 2 3 7 1 4
[12,] 4 1 7 3 2 5 6
[13,] 2 6 7 5 4 3 1
[14,] 4 7 1 2 3 6 5
```

## Checking for Balance

Experimental plans that fit into the row-column scheme may be checked for balance. The function `isGYD` checks whether a given design is a balanced block design with respect to rows, columns or both rows and columns. The user specifies the matrix that represents the design. The design is then checked for balance. If the argument `tables` is set to `TRUE`, matrices giving the number of occurrences of each treatment in each row and column as well as other tables describing the design are given. Continuing the example of the previous section we have

```
> isGYD(design, tables=TRUE)

[1] The design is a generalized latin
square.

$"Number of occurrences of treatments
in d"
```

# Package ‘Crossover’

February 19, 2015

**Type** Package

**Title** Crossover Designs

**Version** 0.1-13

**Author** Kornelius Rohmeyer

**Maintainer** Kornelius Rohmeyer <rohmeier@small-projects.de>

**Description** Package Crossover provides different crossover designs from combinatorial or search algorithms as well as from literature and a GUI to access them.

**Depends** R (>= 3.0.2), rJava (>= 0.8-3), CommonJavaJars (>= 1.0.5),  
JavaGD, ggplot2

**Imports** MASS, crossdes (>= 1.1-1), xtable, methods, Matrix, multcomp,  
stats4, digest

**Suggests** knitr, RUnit, nlme

**SystemRequirements** Java (>= 5.0)

**LinkingTo** Rcpp (>= 0.10.3), RcppArmadillo (>= 0.2.0)

**License** GPL-2

**VignetteBuilder** knitr

**URL** <https://github.com/kornl/Crossover/wiki>

**BugReports** <https://github.com/kornl/Crossover/issues>

**NeedsCompilation** yes

**Repository** CRAN

**Date/Publication** 2014-10-31 23:44:30

## R topics documented:

Crossover-package . . . . .	2
buildSummaryTable . . . . .	2
CrossoverDesign-class . . . . .	3
Crossoverdesigns . . . . .	4
CrossoverGUI . . . . .	6

# Chapter 1

## Introduction

This package provides more than two hundred cross-over design from literature, a search algorithm to find efficient cross-over designs for various models and a graphical user interface (GUI) to find/generate appropriate designs.

The computationally intensive parts of the package, i.e. the search algorithm, is written using the R packages Rcpp and RcppArmadillo (Eddelbuettel and François [2011] and Eddelbuettel and Sanderson [2013]). The GUI is written in Java and uses package rJava (Urbanek [2013]).

### 1.1 Installation

Once it is installed, whenever you start R you can load the Crossover package by entering `library(Crossover)` into the R Console. The graphical user interface as shown in figure 1.1 is started with the command `CrossoverGUI()`.

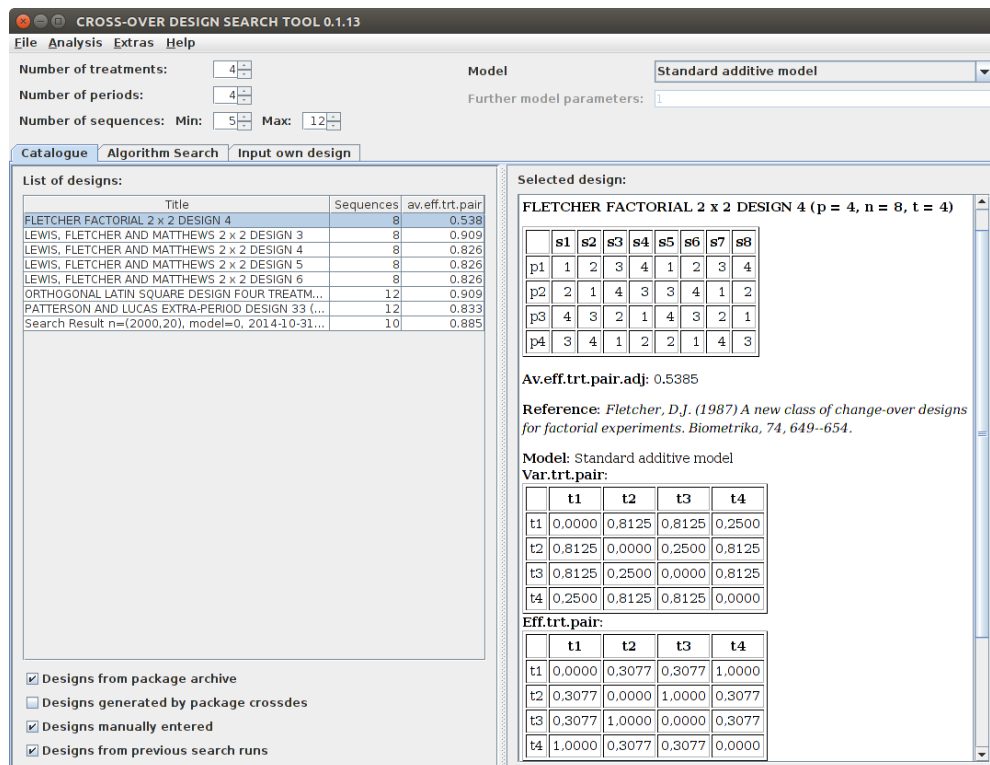


Figure 1.1: Cross-Over Design GUI.

## 1.2.2 Designs in R

Our crossover designs in R are numeric matrices, where the elements represent the treatments, the rows represent the periods and the columns the subjects.

A data frame referencing all available designs and the respective number of treatments, periods and sequences is available by calling the `buildSummaryTable()` function.

Designs referenced in this table can be accessed via the `getDesign()` function. For example the Williams design for three treatments is represented by a  $3 \times 6$ -matrix and assigns each of the three treatments once to each of the six subjects:

```
getDesign("williams3t")

##      s1 s2 s3 s4 s5 s6
## p1  1  2  3  3  1  2
## p2  2  3  1  2  3  1
## p3  3  1  2  1  2  3
## attr(,"reference")
## [1] "Williams, E.J. (1949) Experimental designs balanced for the estimation of residual effects of
## attr(,"signature")
## [1] "p = 3, n = 6, t = 3"
## attr(,"title")
## [1] "WILLIAMS DESIGN THREE TREATMENTS"
```

Each treatment occurs six times, two times in the each period and each treatment follows each other treatment exactly two times.

If we are interested in the variance of the treatment parameter difference estimates, we can use the function `general.carryover`:

```
design <- getDesign("williams3t")
general.carryover(design, model=1)

## $Var.trt.pair
##      [,1] [,2] [,3]
## [1,] 0.0000 0.4167 0.4167
## [2,] 0.4167 0.0000 0.4167
## [3,] 0.4167 0.4167 0.0000
##
## $Var.car.pair
##      [,1] [,2] [,3]
## [1,] 0.00 0.75 0.75
## [2,] 0.75 0.00 0.75
## [3,] 0.75 0.75 0.00
##
## $model
## [1] 1
```

We see that the Williams design is a balanced design.

The following nine models which are discussed in chapter 2 are implemented:

## 3

*When change over time is not linear*

## CONTENTS

3.1	Chapter overview .....	37
3.2	Choosing a functional form .....	38
3.2.1	Function must be adequate for the shape of the data .....	38
3.2.2	Dynamic consistency .....	39
3.2.3	Making predictions: Fits and forecasts .....	43
3.3	Using higher-order polynomials .....	44
3.3.1	Strengths and weaknesses .....	45
3.3.2	Choosing polynomial order .....	46
3.3.3	Orthogonal polynomials .....	47
3.3.4	Interpreting higher-order polynomial effects .....	49
3.4	Example: Word learning .....	51
3.5	Parameter-specific $p$ -values .....	54
3.6	Reporting growth curve analysis results .....	57
3.7	Chapter recap .....	59
3.8	Exercises .....	60

## 3.1 Chapter overview

The previous chapter provided a conceptual overview of growth curve analysis and simple **linear** examples. Of course, time course data in the behavioral, cognitive, and neural sciences are **rarely straight lines**. Typically, the data have complex curved shapes, which means that the **Level 1 model must also have a curved shape**. The choice of the Level 1 model defines a *functional form* for the data; that is, the overall function or shape that will be used to **describe the group and individual data**. This choice is very important because it defines the framework for the whole analysis, so this chapter will describe some options and factors involved in **choosing a functional form with a focus on one particularly good option: *higher-order polynomials***. This approach will be demonstrated with a step-by-step walk through a complete example, including how to estimate parameter-specific  $p$ -values and how to report growth curve analysis results.



### 3.4 Example: Word learning

In Chapter 1, we saw that traditional *t*-test and ANOVA approaches were not effective at capturing the effect of transitional probability (TP) on the rate of novel word learning. These example data are taken from a real experiment (Mirman, Magnuson, Graf Estes, & Dixon, 2008) and reproduced in [Figure 3.7](#). Let's analyze them using **GCA**. The first step should always be to look at the data, both in text form and graphically.

```
> summary(WordLearnEx)
      Subject      TP      Block      Accuracy
244      : 10  Low :280  Min.      : 1.0  Min.      :0.000
253      : 10  High:280 1st Qu.: 3.0  1st Qu.:0.667
302      : 10                      Median : 5.5  Median :0.833
303      : 10                      Mean   : 5.5  Mean   :0.805
305      : 10                      3rd Qu.: 8.0  3rd Qu.:1.000
306      : 10                      Max.   :10.0  Max.   :1.000
(Other):500
```

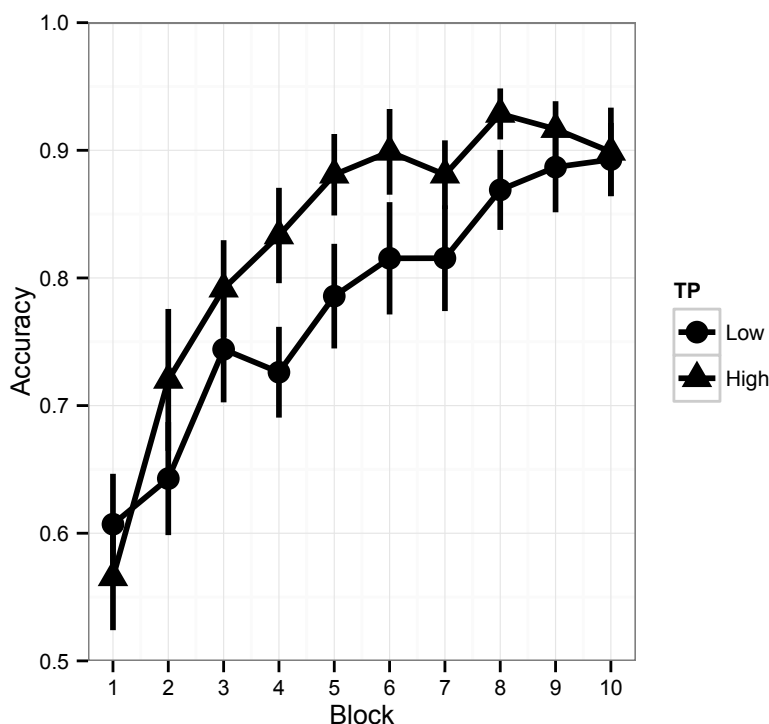
The data frame contains 4 variables:

- **Subject:** A unique identifier for each participant. The identifier is numeric, but treated as a categorical factor. The summary tells us that there are 10 observations per participant.
- **TP:** A categorical between-participants factor with two levels, low and high (within-participants manipulations will be covered in Chapter 4). There are 280 observations in each condition, 10 for each of 28 participants.
- **Block:** A numeric variable indicating training block, ranging from 1 to 10.
- **Accuracy:** Proportion correct for a given participant in a given training block, ranging from 0 to 1.

Here is the code for generating Figure 3.7:

```
> ggplot(WordLearnEx, aes(Block, Accuracy, shape=TP)) +
  stat_summary(fun.y=mean, geom="line", size=1) +
  stat_summary(fun.data=mean_se, geom="pointrange", size=1) +
  theme_bw(base_size=10) +
  coord_cartesian(ylim=c(0.5, 1.0)) +
  scale_x_continuous(breaks=1:10)
```

For data like these, a second-order polynomial should suffice. We'll use orthogonal polynomials for a few reasons. First, in the experiment, participants

**FIGURE 3.7**

Effect of transitional probability (TP) on novel word learning.

learned to match a made-up spoken “word” like *pibu* with a novel geometric shape. All of these “words” were completely novel and arbitrarily paired with shapes and counterbalanced across participants. There were two shape choices on each trial, so it is not very interesting that accuracy would start around 50%, making the y-intercept not very informative. On the other hand, the overall mean accuracy does (partially) reflect faster learning, so the orthogonal intercept will be more informative. Second, orthogonal polynomials will make the linear and quadratic terms uncorrelated, so we will be able to independently evaluate the linear slope and the steepness of the curvature. We can use the `poly` function to create a second-order orthogonal polynomial in the range of `Block`:

```
> t <- poly(unique(WordLearnEx$Block), 2)
```

Now we need to add those orthogonal polynomial values into the original data frame aligned by `Block`. The following command will do that by creating two new variables, `ot1` and `ot2` (for orthogonal time order 1 and orthogonal time

# Formulating (G)lmer models

~~week 6~~  
~~problem 1~~

respiration (dichotomous)  $Pr\{\text{"Good"}\}$

epilepsy (count)  
binomial link  
poisson link

seizure count

Poisson mess  $p(y|\mu) = e^{-\mu} \frac{\mu^y}{y!}$  Link  $\eta = g(\mu) = \log(\mu)$

Bernoulli mess  $p(y|\mu) = \mu^y (1-\mu)^{1-y}$

$\eta = g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$

Respiration ex class handout Sec 13.4 HSAUR

binomial link Level 1  $\eta = \beta_0 + \varepsilon$  flat across months after protocol  
flat trend by (1/subject) in lmer

Level 2  $\underline{\beta_0} = \underline{\gamma_{00}} + \underline{\gamma_{01} \text{ baseline}} + \underline{\gamma_{02} \text{ month}}$   
 $+ \underline{\gamma_{03} \text{ treatment}} + \gamma_{04} \text{ gender}$   
 $+ \gamma_{05} \text{ age} + \gamma_{06} \text{ centre} + u_0$

combined model

substitute  $\beta_0$  to match lmer  
(Level 2 for  $\beta_0$ )  
Rogosa R-session  
posting

Alternative

Level 1 allows trend in logit over months  
 $\eta = \beta_0 + \beta_1 * \text{month} + \varepsilon$  [do month - 1  
compare models so  $\beta_0$  is meaningful]

# Some choices of univariate conditional distributions

## generalized linear mixed models (GLMMs)

- Typical choices of univariate conditional distributions are:
  - ▶ The *Bernoulli* distribution for binary (0/1) data, which has probability mass function

$$p(y|\mu) = \mu^y(1 - \mu)^{1-y}, \quad 0 < \mu < 1, \quad y = 0, 1$$

- ▶ Several independent binary responses can be represented as a *binomial* response, but only if all the Bernoulli distributions have the same mean.
- ▶ The *Poisson* distribution for count (0, 1, ...) data, which has probability mass function

$$p(y|\mu) = e^{-\mu} \frac{\mu^y}{y!}, \quad 0 < \mu, \quad y = 0, 1, 2, \dots$$

- All of these distributions are completely specified by the conditional mean. This is different from the conditional normal (or Gaussian) distribution, which also requires the common scale parameter,  $\sigma$ .

# The canonical link for the Poisson distribution

- The logarithm of the probability mass is

$$\log(p(y|\mu)) = \log(y!) - \mu + y \log(\mu)$$

- Thus, the canonical link function for the Poisson is the *log* link

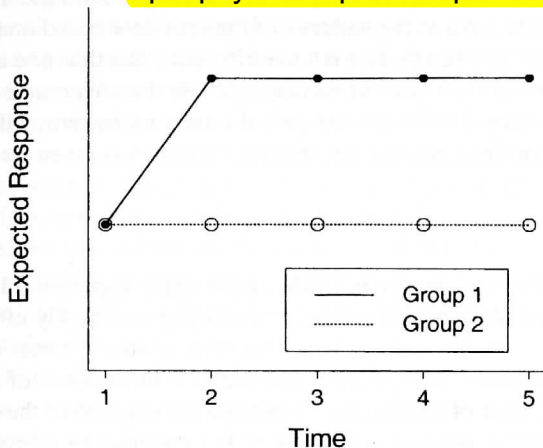
$$\eta = g(\mu) = \log(\mu)$$

- The inverse link is

$$\mu = g^{-1}(\eta) = e^{\eta}$$

## Chap 5 Laird-Ware ALA text

epilepsy or respiration profiles



**Fig. 5.4** Graphical representation of changes in the mean response from baseline (in Group 1) that persist throughout the duration of follow up.



## 5.6 ADJUSTMENT FOR BASELINE RESPONSE

When the data are complete, each of the one-degree-of-freedom tests described in the previous section can be constructed by calculating a univariate summary statistic for each study participant and performing a test for equality of means of these summary statistics in the  $G$  groups. With complete data, group comparisons of these summary statistics are equivalent to applying the corresponding contrast weights to the mean responses. This is because the difference in the means is the mean of the differences when each subject is measured at every occasion. Moreover, for each of the two tests described in detail, mean response minus baseline and AUC minus baseline, the summary statistic corresponds to subtracting the baseline value from a summary of the responses on occasions 2 through  $n$ . For example, for the test for equality of mean response minus baseline, the summary statistic for the  $i^{th}$  participant is given by

$$\frac{(Y_{i2} + Y_{i3} + \cdots + Y_{in})}{n-1} - Y_{i1}. \quad (5.1)$$

With this representation in mind, some analysts have suggested an alternative approach analogous to analysis of covariance (ANCOVA), in which a summary of the response at times 2 through  $n$  becomes the dependent variable and the baseline value enters the analysis as a covariate. When the response variable is the mean at occasions 2 through  $n$  and we wish to test for the equality of the mean in two treatment groups, we can write the corresponding univariate model as

$$Y_i^* = \beta_1 + \beta_2 Y_{i1} + \beta_3 \text{trt}_i + e_i^*, \quad (5.2)$$

where

$$Y_i^* = \frac{(Y_{i2} + Y_{i3} + \cdots + Y_{in})}{n-1}$$

is the mean response at occasions 2 through  $n$  for the  $i^{th}$  subject,  $\text{trt}_i$  is an indicator variable distinguishing the two treatment groups, and  $e_i^*$  is the error term in the univariate model. This model assumes that the data are complete and it cannot be fit with missing data; we defer a discussion of more general approaches for handling baseline response to Section 5.7.

An analysis based on either (5.1) or (5.2) will be especially appealing in settings where initial changes from baseline are expected to persist throughout the duration of follow up. For example, in a trial where the impact of the intervention on changes in the mean response at the start of follow up is expected to be similar to that toward the end of follow up; this pattern for the mean response profiles is illustrated in Figure 5.4. Tests based on (5.1) or (5.2) correspond to a comparison between groups of the mean responses on occasions 2 through  $n$ , with adjustment for baseline, and have  $G-1$  degrees of freedom irrespective of the number of occasions of measurement.

This raises a question about whether one should incorporate the baseline value through the contrast given by (5.1) or through the analysis of covariance model given by (5.2) in a specific application. The answer depends critically on whether the data arose from an observational study or a randomized trial. If the study is an

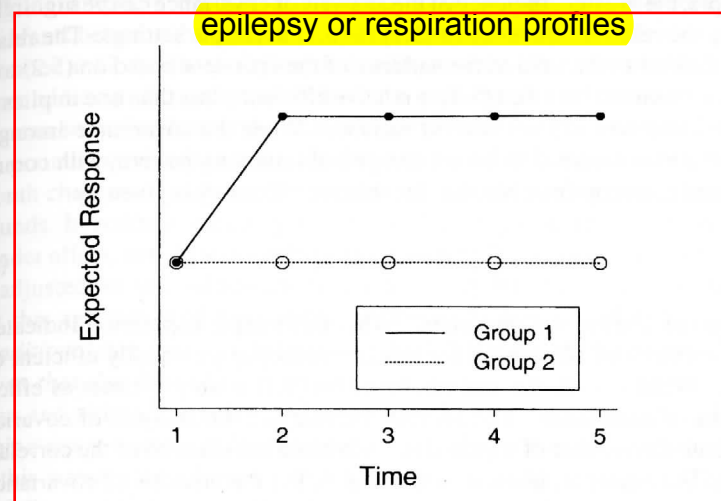


Fig. 5.4 Graphical representation of changes in the mean response from baseline (in Group 1) that persist throughout the duration of follow up.

observational one, for example, a longitudinal study of the determinants of rate of decline of pulmonary function in adults, it is usually not advisable to employ the analysis of covariance approach because the baseline value may be associated with other variables whose effects are to be studied, raising problems of confounding in an analysis intended to describe how the pattern of response over time is influenced by the characteristics of study participants. For example, individuals who are smokers as adults may have smoked during adolescence. If smoking affected the attained pulmonary function level for young adults, then smoking will likely be associated with pulmonary function level later in adult life, even if cigarette smoking does not influence the rate of decline of pulmonary function with age. Thus, adjustment for baseline pulmonary function level using (5.2) could introduce an association between smoking status and rate of decline of pulmonary function, even if the unadjusted rates of decline are nearly equivalent in the various smoking groups.

When participants have been randomized to the several treatment groups and the baseline value has been obtained before any study interventions, adjustment for baseline through analysis of covariance is of interest. In that setting, the mean response at time 1 is independent of treatment assignment. One can then show the one-degree-of-freedom test for equality of response profiles based on a contrast and the corresponding test based on analysis of covariance represent alternative tests of the same null hypothesis and that the test based on the analysis of covariance approach will always be more efficient. That is, the analysis of covariance approach yields estimates of treatment effects with smaller standard errors than those obtained by calculating contrasts.



For example, the greater efficiency of the analysis of covariance can be highlighted by examining the relative efficiency of (5.1) to (5.2) in simple settings. The relative efficiency is defined as the ratio of the variance of the estimator based on (5.2) to the variance of the estimator based on (5.1); a relative efficiency less than one implies that the estimator based on (5.2) has smaller variance. When the covariance among the repeated measures is assumed to have a compound symmetry pattern, with common variance  $\sigma^2$  and common correlation  $\rho$ , the relative efficiency is given by

$$\frac{1}{n} \{1 + (n-1)\rho\}. \quad (5.3)$$

The derivation of (5.3) is not important. What this simple expression indicates is that the two methods of adjustment for baseline response are equally efficient only when  $\rho = 1$ . When  $\rho = 0$ , the analysis based on (5.1) is only  $\frac{1}{n}$  times as efficient as the analysis of covariance. The greater efficiency of the analysis of covariance depends on both the number of repeated measures and the strength of the correlation among them. For example, when  $n = 5$  and  $\rho = 0.4$  the analysis of covariance is approximately twice as efficient as subtracting the baseline response.

In general, the analysis of longitudinal data from a randomized trial is the only setting where we recommend adjustment for baseline through analysis of covariance. In that setting, in contrast to observational studies, adjustment leads to meaningful tests of hypothesis of scientific interest. Moreover, the tests based on the analysis of covariance approach will be more powerful. The notion of adjustment for baseline can also be applied more generally in the analysis of response profiles; in Section 5.7 we compare and contrast a number of alternative strategies for handling the baseline response in more general settings and make recommendations about the preferred strategies in different situations.

We conclude this section by noting that adjustment for baseline in the analysis of longitudinal change is a topic that has generated heated debate among analysts. When longitudinal data arise from an observational study, the two methods of adjusting for baseline described in this section can yield discernibly different and, apparently conflicting, results. This conundrum is also known as *Lord's paradox* (named after Frederic Lord, who eloquently brought the issue to light) and has led many researchers astray over the years. The paradox lies in the interpretation of the two types of analyses and is resolved by noting that these two alternative methods of adjusting for baseline answer qualitatively different scientific questions when the data arise from an observational study. This can be illustrated in the simplest setting where there are two groups or sub-populations (e.g., males and females) measured at two occasions. The overall goal of such a study is to compare the changes in response for the two groups. The analysis that subtracts baseline response, thereby creating a simple change score, addresses the question of whether the two groups differ in terms of their mean change over time. In contrast, adjustment for baseline using analysis of covariance addresses the question of whether an individual belonging to one group is expected to change more (or less) than an individual belonging to the other group, given that they have the same baseline response. The latter question is a conditional one and, depending on the study design, may address a different scientific question than the former.

For example, in an observational study examining gender difference in weight gain of infants between the ages of 12 and 24 months, a measure of body weight might be obtained at 12 months (baseline) and at 24 months. The analysis of the simple change score addresses the question of whether boys and girls differ in terms of their changes in mean body weight over the 12 months of follow up. At baseline, boys are on average  $1\frac{1}{2}$  pounds heavier than girls, but there is no evidence of a gender effect on the 12 month changes in body weight, with boys and girls both gaining approximately  $5\frac{1}{4}$  pounds. In contrast, the analysis of covariance of the same data reveals a discernible gender effect, with boys showing more weight gain than girls. Thus, even though the unadjusted (or *unconditional*) increases in body weight are approximately the same for this age cohort of boys and girls, the analysis of covariance is directed at the *conditional* question of whether boys are expected to gain more weight than girls, given that they have the same initial weight at 12 months. That is, if we compare boys and girls within sub-populations with the same initial weight at 12 months, are their average weights at 24 months the same. When the conditional question is posed in this way we would expect boys to gain more weight than girls. The reasoning is as follows: If a boy and girl have the same initial weight at 12 months then there are two possibilities: (i) the girl is initially overweight and is expected to gain less weight over the 12 months, or (ii) the boy is initially underweight and is expected to gain more.

A more thorough discussion of this issue is beyond the scope of this book, but we advise readers to employ the analysis of covariance approach in longitudinal settings only if the approach and its implications are fully understood.

In summary, the choice between the two methods of adjusting for baseline discussed in this section should be made on substantive grounds. That is, the design of the longitudinal study and the research question of interest should guide the choice of analytic method. The analysis that subtracts baseline response is appropriate when the primary goal of the study is to compare distinct populations in terms of their average change over time. The analysis addresses the question: "Do the populations differ in terms of their average change?" and is appropriate when the data have arisen from either an observational study or a randomized trial. On the other hand, analysis of covariance will, in general, be appropriate only in cases where individuals have been assigned to groups at random (e.g., a randomized trial) or where the population distributions of the baseline responses can reasonably be assumed to be equal (even though the sample means of the baseline responses may differ across groups). In cases where the population distributions of the baseline responses are equal, it is then meaningful to ask the question: "Is the expected change the same in all groups, when we compare individuals having the same baseline response?" Furthermore, the analysis of covariance will provide a more powerful test of group differences. The latter has often been touted as the main reason why analysis of covariance should be the preferred method of adjusting for baseline. This faulty rationale, however, has blinded many researchers to the potential difficulties in interpreting the results of analysis of covariance when the assumption of equal population distributions of baseline response is not tenable. In conclusion, it is the study design and the scientific question of interest,

---

epilepsy*Epilepsy Data*

---

## Description

A randomised clinical trial investigating the effect of an anti-epileptic drug.

## Usage

```
data("epilepsy")
```

## Format

A data frame with 236 observations on the following 6 variables.

treatment the treatment group, a factor with levels placebo and Progabide.

base the number of seizures before the trial.

age the age of the patient.

seizure.rate the number of seizures (response variable).

period treatment period, an ordered factor with levels 1 to 4.

subject the patient ID, a factor with levels 1 to 59.

## Details

In this clinical trial, 59 patients suffering from epilepsy were randomized to groups receiving either the anti-epileptic drug Progabide or a placebo in addition to standard chemotherapy. The numbers of seizures suffered in each of four, two-week periods were recorded for each patient along with a baseline seizure count for the 8 weeks prior to being randomized to treatment and age. The main question of interest is whether taking progabide reduced the number of epileptic seizures compared with placebo.

## Source

P. F. Thall and S. C. Vail (1990), Some covariance models for longitudinal count data with overdispersion. *Biometrics*, **46**, 657–671.

## Examples

```
data("epilepsy", package = "HSAUR2")
library(lattice)
dotplot(I(seizure.rate / base) ~ period | subject, data = epilepsy,
        subset = treatment == "Progabide")
dotplot(I(seizure.rate / base) ~ period | subject, data = epilepsy,
        subset = treatment == "Progabide")
```

## Example 3: Poisson hierarchical model

Breslow and Clayton (1993) analyse data initially provided by Thall and Vail (1990) concerning seizure counts in a randomised trial of anti-convulsant therapy in epilepsy. The table below shows the successive **seizure counts for 59 patients**. Covariates are:

- **treatment (0,1)**
- **8-week baseline seizure counts,**
- **age** in years. The structure of this data is shown below

Patient	y1	y2	y3	y4	Trt	Base	Age
1	5	3	3	3	0	11	31
2	3	5	3	3	0	11	30
3	2	4	0	5	0	6	25
4	4	4	1	4	0	8	36
....							
8	40	20	21	12	0	52	42
9	5	6	6	5	0	12	37
....							
59	1	4	3	2	1	12	37

## EXAMPLE

Clinical trial of anti-epileptic drug Progabide (Thall and Vail, *Biometrics*, 1990)

Randomized, placebo-controlled study of treatment of epileptic seizures with Progabide.

Patients were randomized to treatment with Progabide, or to placebo in addition to standard chemotherapy.

Outcome variable: Count of number of seizures

Measurement schedule: Baseline measurements during 8 weeks prior to randomization. Four measurements during consecutive two-week intervals.

Sample size: 59 epileptics

28 patients on placebo

31 patients on progabide

## Chapter 29

### Count Data: The Epilepsy Study

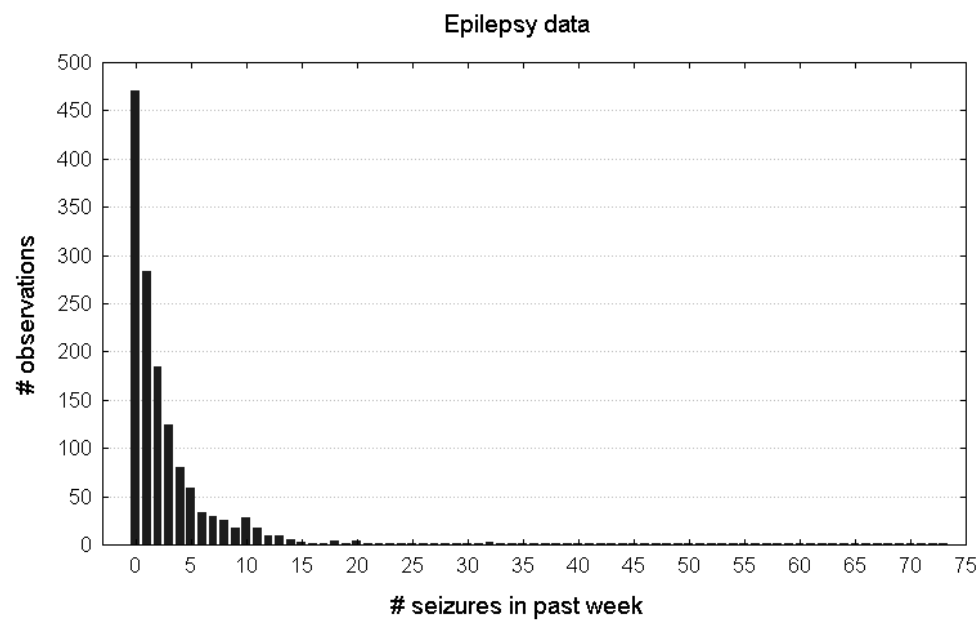
---

- ▷ The epilepsy data
- ▷ Poisson regression
- ▷ Generalized estimating equations
- ▷ Generalized linear mixed models
- ▷ Overview of analyses of the epilepsy study
- ▷ Marginalization of the GLMM

## 29.1 The Epilepsy Data

---

- Consider the epilepsy data:





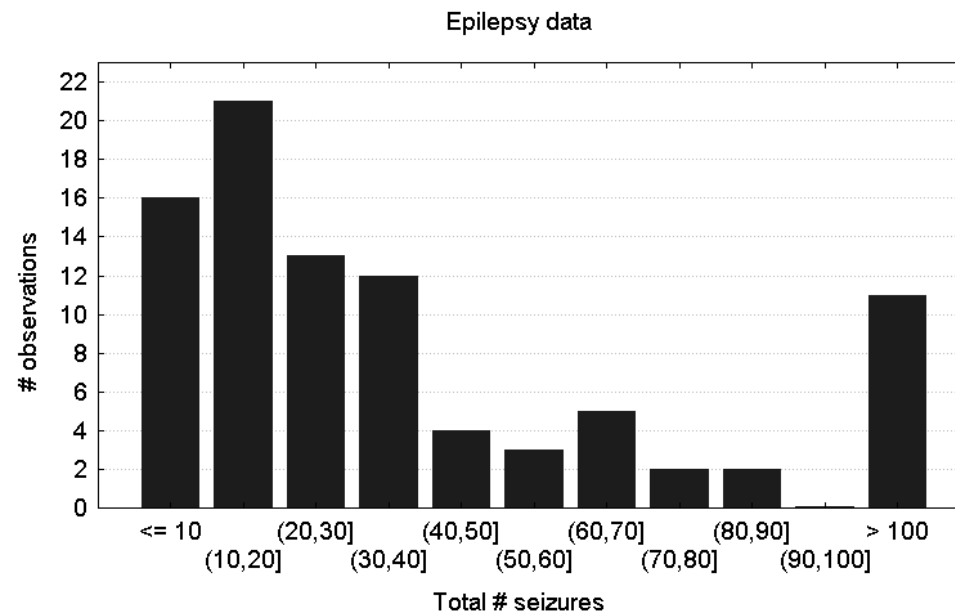
- We want to test for a treatment effect on number of seizures, correcting for the average number of seizures during the 12-week baseline phase, prior to the treatment.
- The response considered now is the total number of seizures a patient experienced, i.e., the sum of all weekly measurements.
- Let  $Y_i$  now be the total number of seizures for subject  $i$ :

$$Y_i = \sum_{j=1}^{n_i} Y_{ij}$$

where  $Y_{ij}$  was the original (longitudinally measured) weekly outcome.



- Histogram:



- As these sums are not taken over an equal number of visits for all subjects, the above histogram is not a 'fair' one as it does not account for differences in  $n_i$  for this.

- We will therefore use the following Poisson model:

$$Y_i \sim \text{Poisson}(\lambda_i)$$

$$\ln(\lambda_i/n_i) = \mathbf{x}_i' \boldsymbol{\beta}$$

n = 2 in epilepsy ex, we make the offset

- Note that the regression model is equivalent to

$$\lambda_i = n_i \exp(\mathbf{x}_i' \boldsymbol{\beta}) = \exp(\mathbf{x}_i' \boldsymbol{\beta} + \ln n_i)$$

- Since  $n_i$  is the number of weeks for which the number of seizures was recorded for subject  $i$ ,  $\exp(\mathbf{x}_i' \boldsymbol{\beta})$  is the average number of seizures per week.
- $\ln n_i$  is called an offset in the above model.
- In our application, the covariates in  $\mathbf{x}_i$  are the treatment as well as the baseline seizure rate.

Residual deviance: 483.22 on 438 degrees of freedom  
AIC: 495.2

HSAUR

Number of Fisher Scoring iterations: 4

---

## Comparing groups

**Figure 13.3** R output of the `summary` method for the `resp_glm` model.

### 13.3.3 Epilepsy

Moving on to the count data in `epilepsy` from Table ??, we begin by calculating the means and variances of the number of seizures for all interactions between treatment and period:

```
R> data("epilepsy", package = "HSAUR2")
R> itp <- interaction(epilepsy$treatment, epilepsy$period)
R> tapply(epilepsy$seizure.rate, itp, mean)

      placebo.1 Progabide.1 placebo.2 Progabide.2 placebo.3
      9.36      8.58      8.29      8.42      8.79
Progabide.3 placebo.4 Progabide.4
      8.13      7.96      6.71

R> tapply(epilepsy$seizure.rate, itp, var)

      placebo.1 Progabide.1 placebo.2 Progabide.2 placebo.3
      102.8      332.7      66.7      140.7      215.3
Progabide.3 placebo.4 Progabide.4
      193.0      58.2      126.9
```

**Figure 13.4** R output of the `summary` method for the `resp_gee1` model (slightly abbreviated).

Some of the variances are considerably larger than the corresponding means, which for a Poisson variable may suggest that overdispersion may be a problem, see Chapter 7.

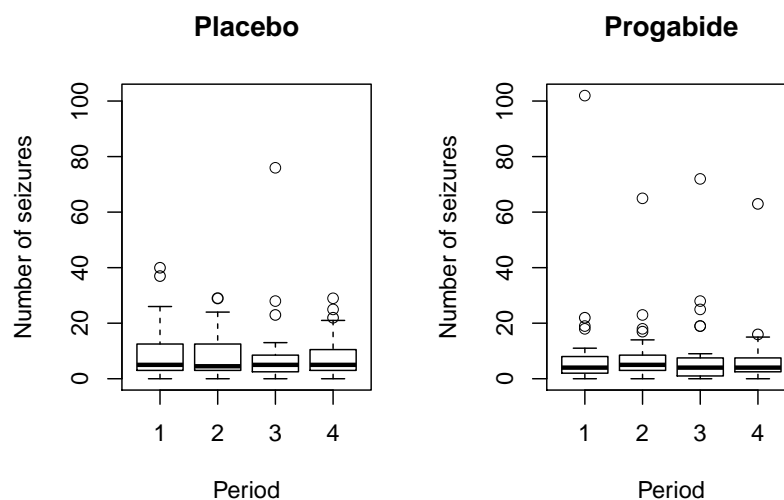
We can now fit a Poisson regression model to the data assuming independence using the `glm` function. We also use the GEE approach to fit an independence structure, followed by an exchangeable structure using the following R code:

```
R> per <- rep(log(2), nrow(epilepsy))
R> epilepsy$period <- as.numeric(epilepsy$period)
R> names(epilepsy)[names(epilepsy) == "treatment"] <- "trt"
R> fm <- seizure.rate ~ base + age + trt + offset(per)
R> epilepsy_glm <- glm(fm, data = epilepsy, family = "poisson")
R> epilepsy_gee1 <- gee(fm, data = epilepsy, family = "poisson",
+   id = subject, corstr = "independence", scale.fix = TRUE,
+   scale.value = 1)
R> epilepsy_gee2 <- gee(fm, data = epilepsy, family = "poisson",
+   id = subject, corstr = "exchangeable", scale.fix = TRUE,
+   scale.value = 1)
R> epilepsy_gee3 <- gee(fm, data = epilepsy, family = "poisson",
+   id = subject, corstr = "exchangeable", scale.fix = FALSE,
+   scale.value = 1)
```

```

R> layout(matrix(1:2, nrow = 1))
R> ylim <- range(epilepsy$seizure.rate)
R> placebo <- subset(epilepsy, treatment == "placebo")
R> progabide <- subset(epilepsy, treatment == "Progabide")
R> boxplot(seizure.rate ~ period, data = placebo,
+         ylab = "Number of seizures",
+         xlab = "Period", ylim = ylim, main = "Placebo")
R> boxplot(seizure.rate ~ period, data = progabide,
+         main = "Progabide", ylab = "Number of seizures",
+         xlab = "Period", ylim = ylim)

```



**Figure 13.6** Boxplots of numbers of seizures in each two-week period post randomisation for placebo and active treatments.

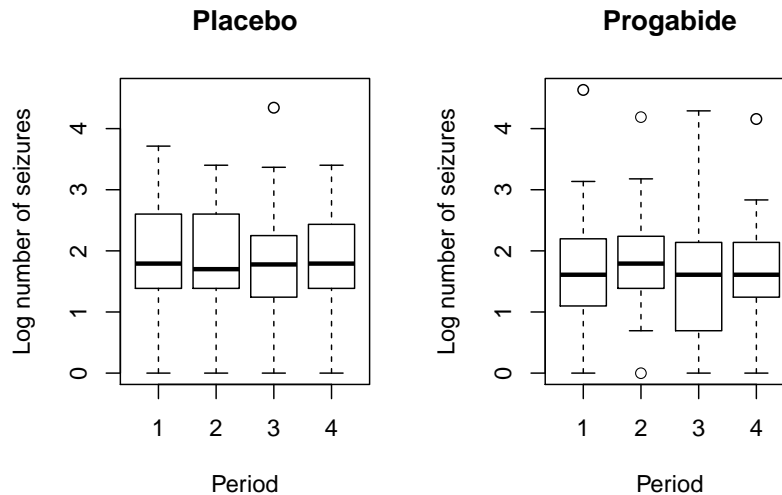
<i>month.C</i>	<i>trttrt</i>	<i>gendermale</i>	<i>age</i>
0.691	8.881	1.227	0.975
<i>centre2</i>			
2.875			

The significance of the effects as estimated by this random effects model and by the GEE model described in Section 13.3.2 is generally similar. But as expected from our previous discussion the estimated coefficients are substantially larger. While the estimated effect of treatment on a randomly sampled individual, given the set of observed covariates, is estimated by the marginal model using GEE to increase the log-odds of being disease free by 1.299, the corresponding estimate from the random effects model is 2.184. These are not inconsistent results but reflect the fact that the models are estimating differ-

```

R> layout(matrix(1:2, nrow = 1))
R> ylim <- range(log(epilepsy$seizure.rate + 1))
R> boxplot(log(seizure.rate + 1) ~ period, data = placebo,
+         main = "Placebo", ylab = "Log number of seizures",
+         xlab = "Period", ylim = ylim)
R> boxplot(log(seizure.rate + 1) ~ period, data = progabide,
+         main = "Progabide", ylab = "Log number of seizures",
+         xlab = "Period", ylim = ylim)

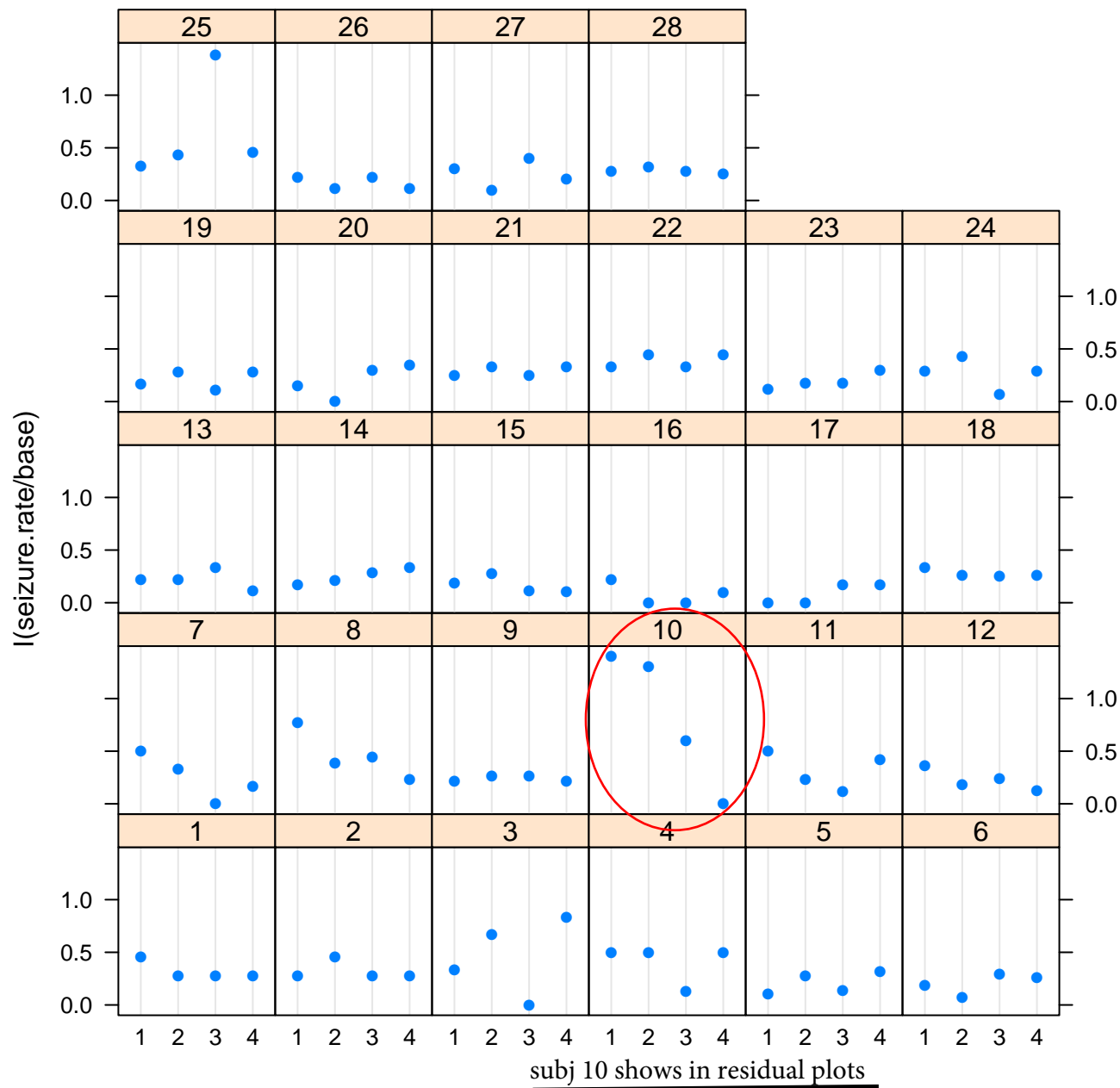
```



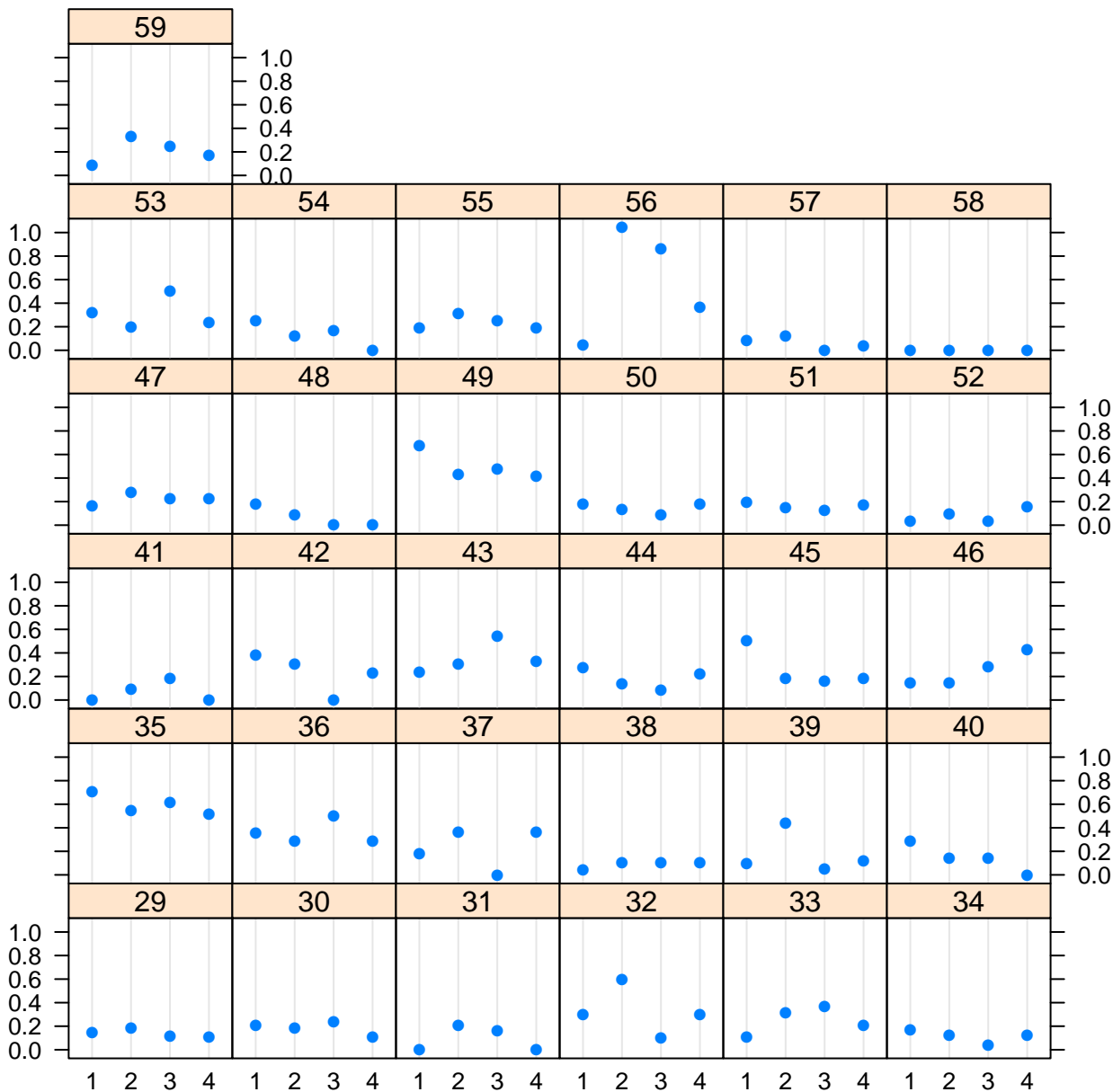
**Figure 13.7** Boxplots of log of numbers of seizures in each two-week period post randomisation for placebo and active treatments.

ent parameters. The random effects estimate is conditional NA in practise. Were we to examine the log-odds of the average predicted probabilities with and without treatment (averaged over the random effects) this would give an estimate comparable to that estimated within the marginal model.

# PLACEBO



# DRUG





AIC: 1732.5

Number of Fisher Scoring iterations: 5

```
> # I didn't do offset so Intercept doesn't match
```

```
> #boxplots of log-seizure
```

```
> placebo<-subset(epilepsy,treatment=="placebo")
```

```
> progabide<-subset(epilepsy,treatment=="Progabide")
```

```
> layout(matrix(1:2,nrow=1))
```

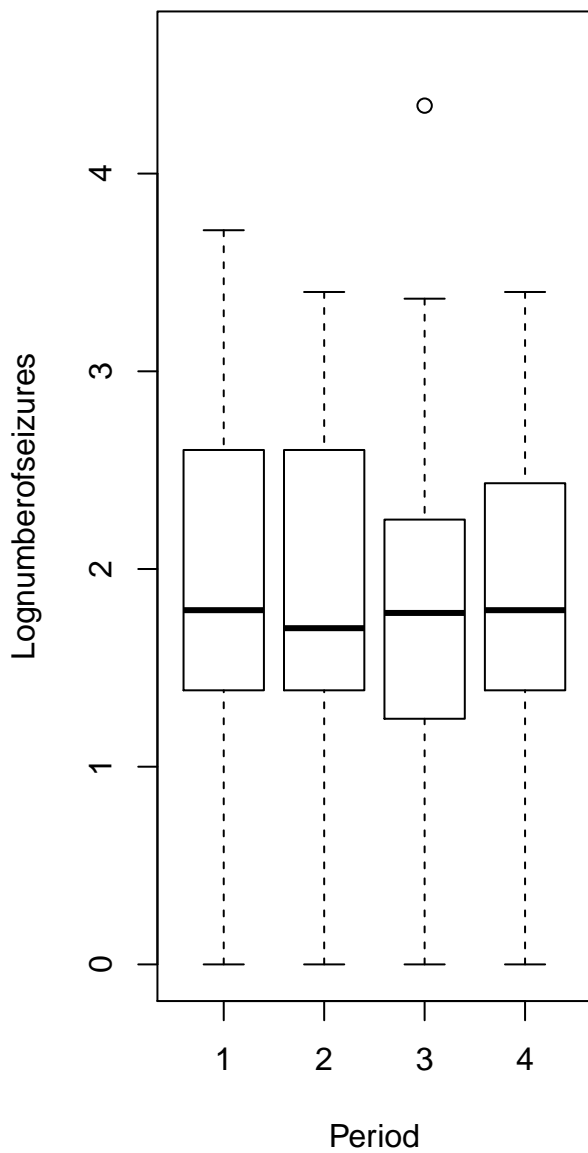
```
> ylim<-range(log(epilepsy$seizure.rate+1))
```

```
> boxplot(log(seizure.rate+1)~period,data=placebo,main="Placebo",ylab="Lognumberofseizu
```

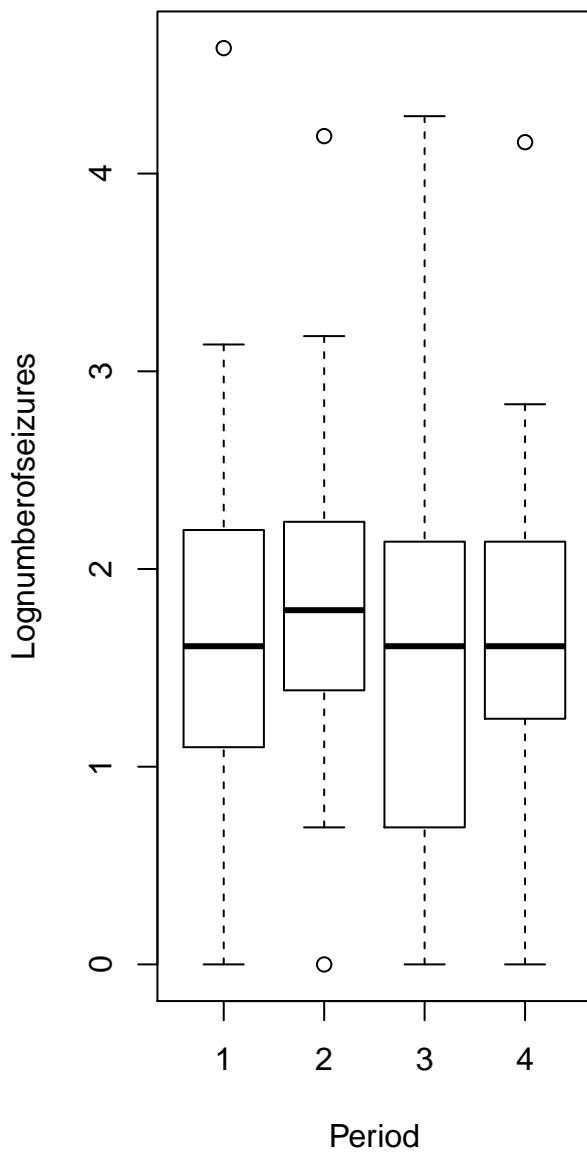
```
> boxplot(log(seizure.rate+1)~period,data=progabide, main="Progabide",ylab="Lognumberof
```

```
>
```

**Placebo**



**Progabide**



---

```
R> summary(epilepsy_glm)
```

**Call:**  
`glm(formula = fm, family = "poisson", data = epilepsy)`

**Deviance Residuals:**

Min	1Q	Median	3Q	Max
-4.436	-1.403	-0.503	0.484	12.322

**Coefficients:**

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.130616	0.135619	-0.96	0.3355
base	0.022652	0.000509	44.48	< 2e-16
age	0.022740	0.004024	5.65	1.6e-08
trtProgabide	-0.152701	0.047805	-3.19	0.0014

---

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 2521.75 on 235 degrees of freedom  
 Residual deviance: 958.46 on 232 degrees of freedom  
 AIC: 1732

Number of Fisher Scoring iterations: 5

---

**Figure 13.8** R output of the `summary` method for the `epilepsy_glm` model.

---

```
R> summary(epilepsy_gee1)
```

...

**Model:**  
 Link: Logarithm  
 Variance to Mean Relation: Poisson  
 Correlation Structure: Independent

...

**Coefficients:**

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	-0.1306	0.135619	-0.963	0.36515	-0.358
base	0.0227	0.000509	44.476	0.00124	18.332
age	0.0227	0.004024	5.651	0.01158	1.964
trtProgabide	-0.1527	0.047805	-3.194	0.17111	-0.892

---

Estimated Scale Parameter: 1

...

---

**Figure 13.9** R output of the `summary` method for the `epilepsy_gee1` model (slightly abbreviated).

see class handout for glm gee; same results

---

```
R> summary(epilepsy_gee2)

...
Model:
  Link:                               Logarithm
Variance to Mean Relation: Poisson
Correlation Structure:      Exchangeable

...

Coefficients:
              Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)   -0.1306   0.200442  -0.652   0.36515  -0.358
base           0.0227   0.000753  30.093   0.00124  18.332
age            0.0227   0.005947   3.824   0.01158   1.964
trtProgabide  -0.1527   0.070655  -2.161   0.17111  -0.892

```

---

```
Estimated Scale Parameter:  1
...
```

---

**Figure 13.10** R output of the `summary` method for the `epilepsy_gee2` model (slightly abbreviated).

---

```
R> summary(epilepsy_gee3)

...
Model:
  Link:                               Logarithm
Variance to Mean Relation: Poisson
Correlation Structure:      Exchangeable

...

Coefficients:
              Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)   -0.1306   0.4522  -0.289   0.36515  -0.358
base           0.0227   0.0017  13.339   0.00124  18.332
age            0.0227   0.0134   1.695   0.01158   1.964
trtProgabide  -0.1527   0.1594  -0.958   0.17111  -0.892

```

---

```
Estimated Scale Parameter:  5.09
...
```

---

**Figure 13.11** R output of the `summary` method for the `epilepsy_gee3` model (slightly abbreviated).

# Epilepsy Clinical Trial

STAT 222  
Week 5

```
> install.packages("HSAUR2")
package 'HSAUR2' successfully unpacked and MD5 sums checked
> library(HSAUR2)
> data(epilepsy)
> dim(epilepsy)
[1] 236 6 > 236/4 [1] 59 > #patients
> library(lattice)
> dotplot(I(seizure.rate/base)~period|subject,data=epilepsy,subset=treatment=="Progabide")
> dotplot(I(seizure.rate/base)~period|subject,data=epilepsy,subset=treatment=="placebo")
> head(epilepsy)
  treatment base age seizure.rate period subject
1    placebo  11  31           5         1      1
110    placebo  11  31           3         2      1
112    placebo  11  31           3         3      1
114    placebo  11  31           3         4      1
2     placebo  11  30           3         1      2
210    placebo  11  30           5         2      2

> itp<-interaction(epilepsy$treatment,epilepsy$period)
> tapply(epilepsy$seizure.rate,itp,mean) #get cell means
placebo.1 Progabide.1 placebo.2 Progabide.2 placebo.3 Progabide.3 placebo.4
9.357143 8.580645 8.285714 8.419355 8.785714 8.129032 7.964286
Progabide.4
6.709677
> # some small advantage for drug
```

```
> #boxplots of log-seizure
> layout(matrix(1:2,nrow=1))
> ylim<-range(log(epilepsy$seizure.rate+1))
> boxplot(log(seizure.rate+1)~period,data=placebo,main="Placebo",ylab="Lognumberofseizures", xlab=
> boxplot(log(seizure.rate+1)~period,data=progabide, main="Progabide",ylab="Lognumberofseizures",
```

```
> per<-rep(log(2),nrow(epilepsy)) # if don't do this offset Intercept doesn't match
> epilepsy_glm <- glm(seizure.rate ~ base + age + treatment + offset(per), data = epilepsy,
family = "poisson")
```

```
> summary(epilepsy_glm)
Call:
glm(formula = seizure.rate ~ base + age + treatment + offset(per),
family = "poisson", data = epilepsy)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.4360 -1.4034 -0.5029  0.4842 12.3223
Coefficients:
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.1306156	0.1356191	-0.963	0.3355
base	0.0226517	0.0005093	44.476	< 2e-16 ***
age	0.0227401	0.0040240	5.651	1.59e-08 ***
treatmentProgabide	-0.1527009	0.0478051	-3.194	0.0014 **

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)
Null deviance: 2521.75 on 235 degrees of freedom
Residual deviance: 958.46 on 232 degrees of freedom
AIC: 1732.5
```

```
# if don't use covariates, then can't find sig treatment effect
> epilepsy_glm2 <- glm(seizure.rate ~ treatment + offset(per), data = epilepsy, family = "poisson")
> summary(epilepsy_glm2)
Call: glm(formula = seizure.rate ~ treatment + offset(per), family = "poisson",
data = epilepsy)
Coefficients:
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1.45841	0.03222	45.258	<2e-16 ***
treatmentProgabide	-0.07717	0.04529	-1.704	0.0884 .

```
> install.packages("gee")
```

Thy

Generalized Estimating Eqs gee  
Liang-Zeger

see HSAUR chap for plots  
and code

glm model  
ignores indiv  
trajectories linked,  
conditions on  
covariates

# Package ‘gee’

March 15, 2012

**Title** Generalized Estimation Equation solver

**Version** 4.13-18

**Depends** stats

**Suggests** MASS

**Date** 2012-03-14

**DateNote** Gee version 1998-01-27

**Author** Vincent J Carey. Ported to R by Thomas Lumley (versions 3.13 and 4.4) and Brian Ripley <ripley@stats.ox.ac.uk> (version 4.13). Note that maintainers are not available to give advice on using a package they did not author.

**Maintainer** Brian Ripley <ripley@stats.ox.ac.uk>

**Description** Generalized Estimation Equation solver

**License** GPL-2

**Repository** CRAN

**Date/Publication** 2012-03-15 07:28:20

## R topics documented:

gee . . . . .	2
Index	5

## References

- Liang, K.Y. and Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73** 13–22.
- Zeger, S.L. and Liang, K.Y. (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42** 121–130.

## B. SEMIPARAMETRIC REGRESSION USING GEE

First introduced by Liang and Zeger (1986); see also Diggle, Liang and Zeger, (1994). Instead of attempting to model the within-subject covariance structure, treat it as a nuisance and simply model the mean response.

$$\begin{aligned}y_i &= (y_{i1}, y_{i2}, \dots, y_{i,n_i})^T \\y_{ij} &= \text{discrete or continuous response} \\E(y_{ij}) &= \mu_{ij}; \text{ mean response} \\g(\mu_i) &= X_i\beta \quad \text{link function} \\\text{Cov}(y_i) &= \Delta_i^{1/2} R_i(\alpha) \Delta_i^{1/2}\end{aligned}$$

where  $R_i$  is a ‘working correlation matrix’ representing a guess at the true correlation structure.



**gee*****Function to solve a Generalized Estimation Equation Model*****Description**

Produces an object of class "gee" which is a Generalized Estimation Equation fit of the data.

**Usage**

```
gee(formula, id,
    data, subset, na.action,
    R = NULL, b = NULL,
    tol = 0.001, maxiter = 25,
    family = gaussian, corstr = "independence",
    Mv = 1, silent = TRUE, contrasts = NULL,
    scale.fix = FALSE, scale.value = 1, v4.4compat = FALSE)
```

**Arguments**

formula	a formula expression as for other regression models, of the form response ~ predictors. See the documentation of <a href="#">lm</a> and <a href="#">formula</a> for details.
id	a vector which identifies the clusters. The length of id should be the same as the number of observations. Data are assumed to be sorted so that observations on a cluster are contiguous rows for all entities in the formula.
data	an optional data frame in which to interpret the variables occurring in the formula, along with the id and n variables.
subset	expression saying which subset of the rows of the data should be used in the fit. This can be a logical vector (which is replicated to have length equal to the number of observations), 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.
na.action	a function to filter missing data. For gee only <code>na.omit</code> should be used here.
R	a square matrix of dimension maximum cluster size containing the user specified correlation. This is only appropriate if <code>corstr = "fixed"</code> .
b	an initial estimate for the parameters.
tol	the tolerance used in the fitting algorithm.
maxiter	the maximum number of iterations.
family	a family object: a list of functions and expressions for defining link and variance functions. Families supported in gee are gaussian, binomial, poisson, Gamma, and quasi; see the <a href="#">glm</a> and <a href="#">family</a> documentation. Some links are not currently available: $1/\mu^2$ and $\sqrt{\cdot}$ have not been hard-coded in the 'cgee' engine at present. The inverse gaussian variance function is not available. All combinations of remaining functions can be obtained either by family selection or by the use of quasi.

```
> library(gee)
```

```
> fm <- seizure.rate ~ base + age + trt + offset(per)
> epilepsy_gee1 <- gee(fm, data = epilepsy, family = "poisson",
+ id = subject, corstr = "independence", scale.fix = TRUE,
+ scale.value = 1)
```

```
Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
running glm to get initial regression estimate
```

(Intercept)	base	age	treatmentProgabide
-0.13061561	0.02265174	0.02274013	-0.15270095

```
> summary(epilepsy_gee1)
```

```
GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
gee S-function, version 4.13 modified 98/01/27 (1998)
```

```
Model:
```

```
Link: Logarithm
Variance to Mean Relation: Poisson
Correlation Structure: Independent
```

```
Call:
```

```
gee(formula = fm, id = subject, data = epilepsy, family = "poisson",
corstr = "independence", scale.fix = TRUE, scale.value = 1)
```

```
Summary of Residuals:
```

Min	1Q	Median	3Q	Max
-4.9195387	0.1808059	1.7073405	4.8850644	69.9658560

```
Coefficients:
```

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	-0.13061561	0.1356191185	-0.9631062	0.365148155	-0.3577058
base	0.02265174	0.0005093011	44.4761250	0.001235664	18.3316325
age	0.02274013	0.0040239970	5.6511312	0.011580405	1.9636736
treatmentProgabide	-0.15270095	0.0478051054	-3.1942393	0.171108915	-0.8924196

"gee"

not sig

```
> epilepsy_lmer <- lmer(seizure.rate ~ base + age + treatment + offset(per) + (period|subject),
data = epilepsy, family = "poisson")
```

must do glmer

```
> summary(epilepsy_lmer)
```

```
Generalized linear mixed model fit by the Laplace approximation
```

```
Formula: seizure.rate ~ base + age + treatment + offset(per) + (period | subject)
```

```
Data: epilepsy
```

```
AIC BIC logLik deviance
```

```
505.5 554 -238.8 477.5
```

```
Random effects:
```

Groups	Name	Variance	Std.Dev.	Corr
subject	(Intercept)	0.27139	0.52095	
	period.L	0.11981	0.34613	0.034
	period.Q	0.14247	0.37746	-0.494 -0.593
	period.C	0.11981	0.34613	-0.274 -0.178 0.098

```
Number of obs: 236, groups: subject, 59
```

```
Fixed effects:
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.215683	0.373725	-0.577	0.5639
base	0.026197	0.002374	11.035	<2e-16 ***
age	0.016859	0.011476	1.469	0.1418
treatmentProgabide	-0.306282	0.140690	-2.177	0.0295 *

```
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Correlation of Fixed Effects:
```

	(Intr)	base	age
base	-0.401		
age	-0.939	0.201	
trtmntPrghd	-0.304	-0.002	0.130

```
> exp(-.1527) > exp(-.3063)
[1] 0.8583872 [1] 0.7361657
```

```
# Laird-Ware ALA p.349 "patient treated with progabide the expected decrease in seizures
is approx 26% (exp(-.3069) ~ .74)"
```

Ch12 bl/lan Laird uses diff model  
period effects, no covariates  
same result as

try lmer  
"poisson"  
no time  
trend, just  
level effect  
of treatment  
(1/subject)

CI by exp(logconfint())

```
> install.packages("gee")
> library(gee)
```

```
> fm <- seizure.rate ~ base + age + trt + offset(per)
> epilepsy_geel <- gee(fm, data = epilepsy, family = "poisson",
+ id = subject, corstr = "independence", scale.fix = TRUE,
+ scale.value = 1)
Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
running glm to get initial regression estimate
```

	(Intercept)	base	age	treatment	Progabide
	-0.13061561	0.02265174	0.02274013		-0.15270095

```
> summary(epilepsy_geel)
GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
gee S-function, version 4.13 modified 98/01/27 (1998)
Model:
Link: Logarithm
Variance to Mean Relation: Poisson
Correlation Structure: Independent
Call:
gee(formula = fm, id = subject, data = epilepsy, family = "poisson",
    corstr = "independence", scale.fix = TRUE, scale.value = 1)
```

```
Summary of Residuals:
    Min       1Q   Median       3Q      Max
-4.9195387  0.1808059  1.7073405  4.8850644  69.9658560

Coefficients:
              Estimate   Naive S.E.   Naive z Robust S.E.   Robust z
(Intercept)   -0.13061561  0.1356191185 -0.9631062  0.365148155 -0.3577058
base           0.02265174  0.0005093011  44.4761250  0.001235664  18.3316325
age            0.02274013  0.0040239970   5.6511312  0.011580405   1.9636736
treatmentProgabide -0.15270095  0.0478051054 -3.1942393  0.171108915 -0.8924196
```

```
> epilepsy_lmer <- lmer(seizure.rate ~ base + age + treatment + offset(per) +(period|subject),
+ data = epilepsy, family = "poisson")
```

alternative (1|subject) no trend

```
> summary(epilepsy_lmer)
Generalized linear mixed model fit by the Laplace approximation
Formula: seizure.rate ~ base + age + treatment + offset(per) + (period | subject)
Data: epilepsy
AIC BIC logLik deviance
505.5 554 -238.8 477.5
Random effects:
Groups Name Variance Std.Dev. Corr
subject (Intercept) 0.27139 0.52095
period.L 0.11981 0.34613 0.034
period.Q 0.14247 0.37746 -0.494 -0.593
period.C 0.11981 0.34613 -0.274 -0.178 0.098
Number of obs: 236, groups: subject, 59
Fixed effects:
```

Ordered Factor, see str()

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.215683	0.373725	-0.577	0.5639
base	0.026197	0.002374	11.035	<2e-16 ***
age	0.016859	0.011476	1.469	0.1418
treatmentProgabide	-0.306282	0.140690	-2.177	0.0295 *

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Correlation of Fixed Effects:
(Intr) base age
base      -0.401
age       -0.939  0.201
trtmntPrghbd -0.304 -0.002  0.130
```

```
> exp(-.1527)      > exp(-.3063)
[1] 0.8583872      [1] 0.7361657
# Laird-Ware ALA p.349 "patient treated with progabide the expected decrease in seizures
is approx 26% (exp(-.3069) ~ .74)"
```

```

Epilepsy Ex Comparing lmer models HW5 Prob 5b
R version 2.15.2 (2012-10-26) -- "Trick or Treat"
> library(HSAUR2)
package 'HSAUR2' was built under R version 2.15.3
> data(epilepsy)
> str(epilepsy)
'data.frame': 236 obs. of 6 variables:
 $ treatment : Factor w/ 2 levels "placebo","Progabide": 1 1 1 1 1 1 1 1 1 1 ...
 $ base      : int 11 11 11 11 11 11 11 11 6 6 ...
 $ age       : int 31 31 31 31 30 30 30 30 25 25 ...
 $ seizure.rate: int 5 3 3 3 3 5 3 3 2 4 ...
 $ period    : Ord.factor w/ 4 levels "1"<"2"<"3"<"4": 1 2 3 4 1 2 3 4 1 2 ...
 $ subject   : Factor w/ 59 levels "1","2","3","4",...: 1 1 1 1 2 2 2 2 3 3 ...

> library(lme4)
> per<-rep(log(2),nrow(epilepsy)) # if don't do this offset Intercept doesn't match
> ep0 <- lmer(seizure.rate ~ treatment + offset(per) +(1|subject), data = epilepsy, family = "p")
> ep1 <- lmer(seizure.rate ~ treatment + offset(per) +(period|subject), data = epilepsy, family = "p")
> anova(ep0, ep1)
Data: epilepsy
Models:
ep0: seizure.rate ~ treatment + offset(per) + (1 | subject)
ep1: seizure.rate ~ treatment + offset(per) + (period | subject)
    Df    AIC    BIC logLik Chisq Chi Df Pr(>Chisq)
ep0  3 641.20 651.59 -317.60
ep1 12 561.69 603.25 -268.84 97.516      9 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(ep1)
Generalized linear mixed model fit by the Laplace approximation
Formula: seizure.rate ~ treatment + offset(per) + (period | subject)
Data: epilepsy
    AIC    BIC logLik deviance
561.7 603.3 -268.8 537.7
Random effects:
Groups Name Variance Std.Dev. Corr
subject (Intercept) 0.86624 0.93072
        period.L    0.10478 0.32370 -0.138
        period.Q    0.14632 0.38252 -0.164 -0.662
        period.C    0.12144 0.34848 -0.247 -0.255 0.149
Number of obs: 236, groups: subject, 59

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.0703     0.1758   6.088 1.14e-09 ***
treatmentProgabide -0.3187     0.2445  -1.303   0.192
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
      (Intr)
trtmntPrgrbd -0.719

> ep2 <- lmer(seizure.rate ~ treatment + offset(per) + base +(1|subject), data = epilepsy, fami
> ep3 <- lmer(seizure.rate ~ treatment + offset(per) + base +(period|subject), data = epilepsy,
> anova(ep2, ep3)
Data: epilepsy
Models:
ep2: seizure.rate ~ treatment + offset(per) + base + (1 | subject)
ep3: seizure.rate ~ treatment + offset(per) + base + (period | subject)
    Df    AIC    BIC logLik Chisq Chi Df Pr(>Chisq)
ep2  4 588.89 602.74 -290.44
ep3 13 505.43 550.46 -239.71 101.46      9 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(ep3)

```

```

Generalized linear mixed model fit by the Laplace approximation
Formula: seizure.rate ~ treatment + offset(per) + base + (period | subject)
Data: epilepsy
AIC BIC logLik deviance
505.4 550.5 -239.7 479.4
Random effects:
Groups Name Variance Std.Dev. Corr
subject (Intercept) 0.28151 0.53058
period.L 0.11940 0.34554 -0.025
period.Q 0.14149 0.37615 -0.432 -0.592
period.C 0.11829 0.34393 -0.304 -0.181 0.098
Number of obs: 236, groups: subject, 59

```

```

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.296535   0.130903   2.265   0.0235 *
treatmentProgabide -0.328494   0.142636  -2.303   0.0213 *
base         0.025390   0.002388  10.632 <2e-16 ***
---

```

```

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

Correlation of Fixed Effects:

```

```

(Intr) trtmnP
trtmnPrgbd -0.534
base       -0.631 -0.030

```

```

> ep4 <- lmer(seizure.rate ~ base + age + treatment + offset(per) + (1|subject), data = epilepsy, f
> ep5 <- lmer(seizure.rate ~ base + age + treatment + offset(per) + (period|subject), data = epilep
> anova(ep4, ep5)

```

```

Data: epilepsy
Models:
ep4: seizure.rate ~ base + age + treatment + offset(per) + (1 | subject)
ep5: seizure.rate ~ base + age + treatment + offset(per) + (period |
ep5: subject)
      Df    AIC    BIC logLik Chisq Chi Df Pr(>Chisq)
ep4   5 589.63 606.95 -289.81
ep5  14 505.51 554.01 -238.76 102.11      9 < 2.2e-16 ***
---

```

```

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

> summary(ep5)

```

```

Generalized linear mixed model fit by the Laplace approximation
Formula: seizure.rate ~ base + age + treatment + offset(per) + (period | subject)
Data: epilepsy
AIC BIC logLik deviance
505.5 554 -238.8 477.5
Random effects:
Groups Name Variance Std.Dev. Corr
subject (Intercept) 0.27139 0.52095
period.L 0.11981 0.34614 0.034
period.Q 0.14248 0.37746 -0.494 -0.593
period.C 0.11981 0.34613 -0.274 -0.178 0.098
Number of obs: 236, groups: subject, 59

```

```

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.215682   0.373724  -0.577   0.5639
base         0.026197   0.002374  11.035 <2e-16 ***
age          0.016859   0.011476   1.469   0.1418
treatmentProgabide -0.306274   0.140690  -2.177   0.0295 *
---

```

```

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

Correlation of Fixed Effects:

```

```

(Intr) base age
base    -0.401
age     -0.939 0.201

```

```

trtmntPrgbd -0.304 -0.002 0.130
> # age doesn't help much as an additional covariate; period helps with any set of covariates
> anova(ep3, ep5)
Data: epilepsy
Models:
ep3: seizure.rate ~ treatment + offset(per) + base + (period | subject)
ep5: seizure.rate ~ base + age + treatment + offset(per) + (period |
ep5:      subject)
      Df      AIC      BIC  logLik  Chisq Chi Df Pr(>Chisq)
ep3 13 505.43 550.46 -239.71
ep5 14 505.51 554.01 -238.76 1.9143      1 0.1665
> #so I would go with ep3
> # alternative to base as covariate would be outcome seizure/base a bit messier
>

```

```

> exp(-.3285) [1] 0.7200029 > #maybe 28% decrease in seizures better answer than 26%

```

Note: lmlist does glm;

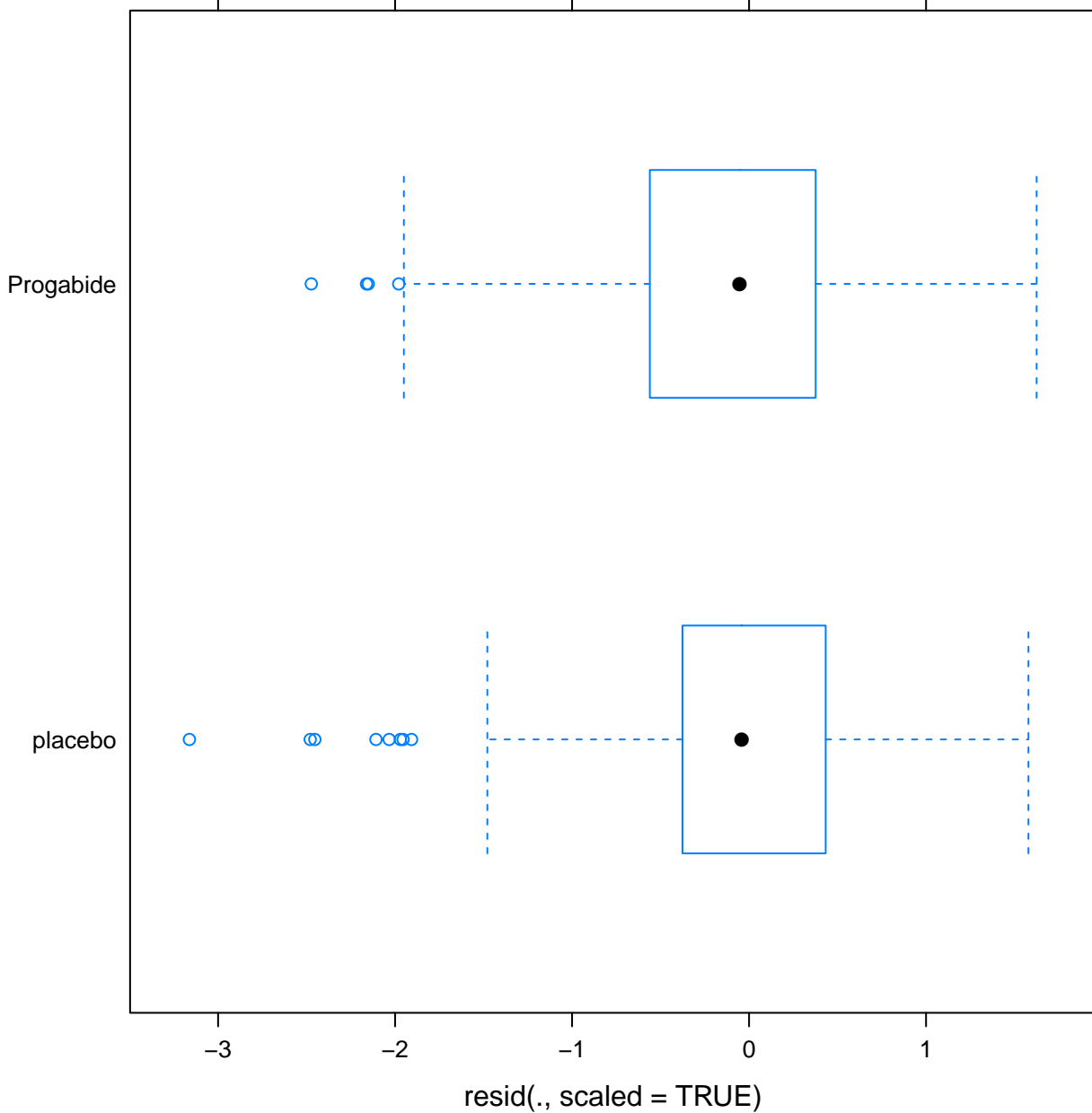
```
fm2 <- lmlist(y2 ~ 1 | g, data=d, family=binomial)
```

```

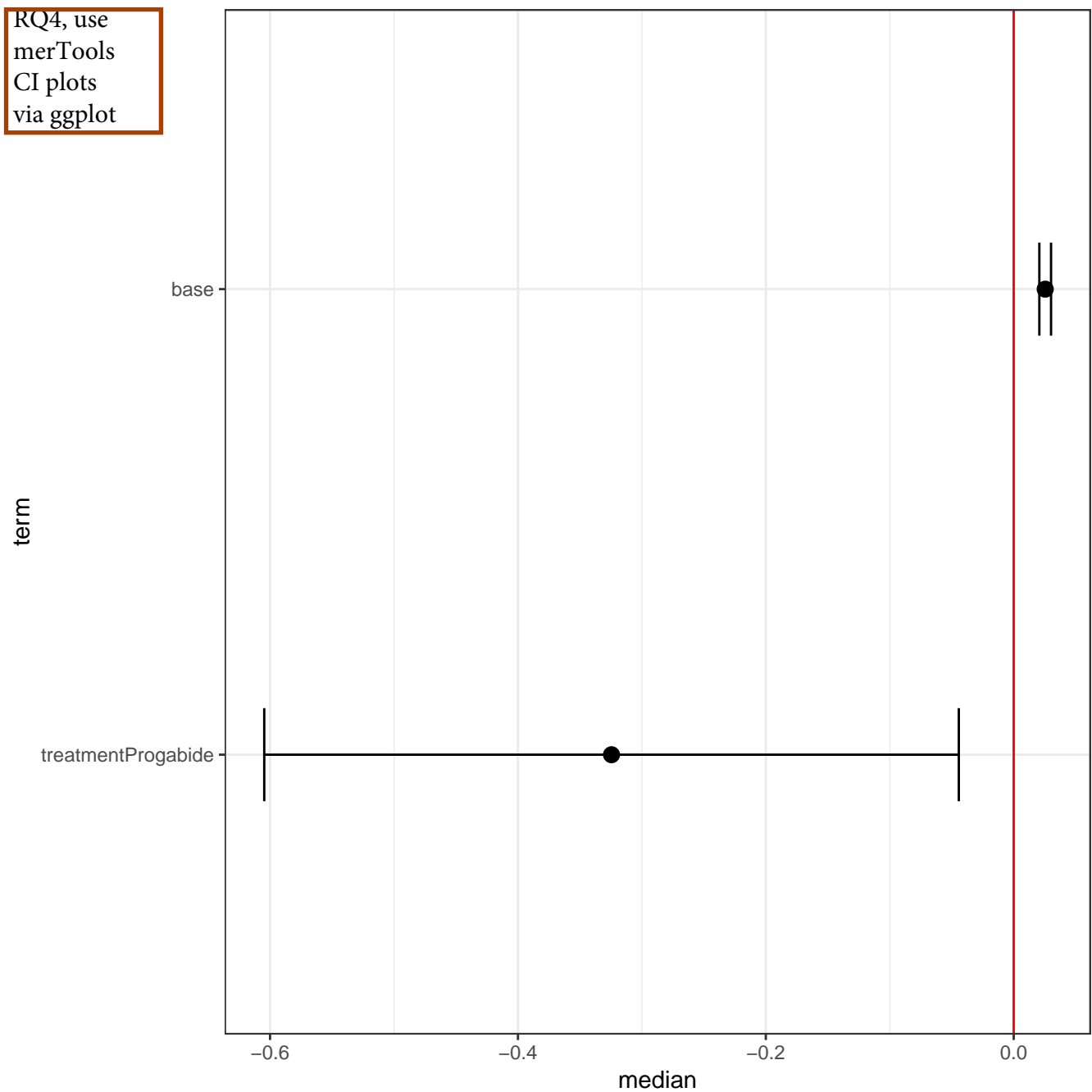
> exp(fixef(ep3))
(Intercept) treatmentProgabide      base
1.3451900      0.7200072      1.0257155

```

From RQ4, assorted residual plots for ep3



RQ4, use  
merTools  
CI plots  
via ggplot





# Formulating (G)lmer models

respiration (dichotomous)  $Pr\{\text{"Good"}\}$   
 binomial link  
epilepsy (count) seizure count  
 poisson link

Poisson mess  $p(y|\mu) = e^{-\mu} \frac{\mu^y}{y!}$  Link  
 $\eta = g(\mu) = \log(\mu)$

Bernoulli mess  $p(y|\mu) = \mu^y (1-\mu)^{1-y}$   
 $\eta = g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$

Respiration ex class handout Sec 13.4 HSAUR

binomial link  
 Level 1  $\eta = \beta_0 + \epsilon$  flat across months after protocol  
 flat trend by (1/subject) in lmer  
 Level 2  $\beta_0 = \gamma_{00} + \gamma_{01} \text{ baseline} + \gamma_{02} \text{ month}$   
 $+ \gamma_{03} \text{ treatment} + \gamma_{04} \text{ gender}$   
 $+ \gamma_{05} \text{ age} + \gamma_{06} \text{ centre} + u_0$

combined model  
 substitute  $\beta_0$  to match lmer  
 (Level 2 for  $\beta_0$ )  
 Rogers R-session  
 posting

Alternative

Level 1 allows trend in logit over months  
 $\eta = \beta_0 + \beta_1 * \text{month} + \epsilon$  [do month-1  
 compare models so  $\beta_0$  is meaningful

## 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 <code>~</code> operator and the terms, separated by <code>+</code> operators, on the right. Random-effects terms are distinguished by vertical bars ( <code>" "</code> ) 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 <code>lmer</code> is called. While data is optional, the package authors <i>strongly</i> recommend its use, especially when later applying methods such as <code>update</code> and <code>drop1</code> to the fitted model ( <i>such methods are not guaranteed to work properly if data is omitted</i> ). 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 <a href="#">glm</a> and <a href="#">family</a> .  |
| control               | a list (of correct class, <a href="#">resulting from lmerControl()</a> or <a href="#">glmerControl()</a> respectively) containing control parameters, including the nonlinear optimizer to be used and parameters to be passed through to the nonlinear optimizer, see the <code>*lmerControl</code> documentation for details.   |
| <a href="#">start</a> | a named list of starting values for the parameters in the model, or a numeric vector. A numeric <code>start</code> argument will be used as the starting value of theta. If <code>start</code> is a list, the <code>theta</code> 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 <code>start[["fixef"]]</code> , are used as starting values for the   |

	second optimization step. If <code>start</code> has both <code>fixef</code> and <code>theta</code> elements, the first optimization step is skipped. For more details or finer control of optimization, see <a href="#">modular</a> .
<code>verbose</code>	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.
<code>nAGQ</code>	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 <a href="#">Details</a> .)
<code>subset</code>	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.
<code>weights</code>	an optional vector of ‘prior weights’ to be used in the fitting process. Should be <code>NULL</code> or a numeric vector.
<code>na.action</code>	a function that indicates what should happen when the data contain NAs. The default action ( <code>na.omit</code> , inherited from the ‘factory fresh’ value of <code>getOption("na.action")</code> ) strips any observations with any missing values in any variables.
<code>offset</code>	this can be used to specify an <i>a priori</i> known component to be included in the linear predictor during fitting. This should be <code>NULL</code> or a numeric vector of length equal to the number of cases. One or more <a href="#">offset</a> terms can be included in the formula instead or as well, and if more than one is specified their sum is used. See <a href="#">model.offset</a> .
<code>contrasts</code>	an optional list. See the <code>contrasts.arg</code> of <a href="#">model.matrix.default</a> .
<code>mustart</code>	optional starting values on the scale of the conditional mean, as in <a href="#">glm</a> ; see there for details.
<code>etastart</code>	optional starting values on the scale of the unbounded predictor as in <a href="#">glm</a> ; see there for details.
<code>devFunOnly</code>	logical - return only the deviance evaluation function. Note that because the deviance function operates on variables stored in its environment, it may not return <i>exactly</i> the same values on subsequent calls (but the results should always be within machine tolerance).
<code>...</code>	other potential arguments. A <code>method</code> argument was used in earlier versions of the package. Its functionality has been replaced by the <code>nAGQ</code> 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
```

```
## these data are notoriously difficult to fit
## Not run:
if (require("HSAUR2")) {
  gm2 <- glmer(outcome~treatment*visit+(1|patientID),
               data=toenail,
               family=binomial,nAGQ=20)
}

## End(Not run)
```

**glmer.nb***Fitting Negative Binomial GLMMs***Description**

Fits a generalized linear mixed-effects model (GLMM) for the negative binomial family, building on [glmer](#), and initializing via [theta.ml](#) from **MASS**.

**Usage**

```
glmer.nb(..., interval = log(th) + c(-3, 3),
          tol = 5e-5, verbose = FALSE, nb.control = NULL,
          initCtrl = list(limit = 20, eps = 2*tol, trace = verbose,
                           theta = NULL))
```

**Arguments**

...	arguments as for <code>glmer(.)</code> such as formula, data, control, etc, but <i>not</i> family!
interval	interval in which to start the optimization. The default is symmetric on log scale around the initially estimated theta.
tol	tolerance for the optimization via <a href="#">optimize</a> .
verbose	<a href="#">logical</a> indicating how much progress information should be printed during the optimization. Use <code>verbose = 2</code> (or larger) to enable <code>verbose=TRUE</code> in the <a href="#">glmer()</a> calls.
nb.control	optional <a href="#">list</a> , like <code>glmerControl()</code> , used in <code>refit(*, control = control.nb)</code> during the optimization.
initCtrl	( <i>experimental, do not rely on this:</i> ) a <a href="#">list</a> with named components as in the default, passed to <a href="#">theta.ml</a> (package <b>MASS</b> ) for the initial value of the negative binomial parameter theta. May also include a theta component, in which case the initial estimation step is skipped

**Value**

An object of class `glmerMod`, for which many methods are available (e.g. `methods(class="glmerMod")`), see [glmer](#).

---

respiratory

---

*Respiratory Illness Data*

### **Description**

The respiratory status of patients recruited for a randomised clinical multicenter trial.

### **Usage**

```
data("respiratory")
```

## Format

A data frame with 555 observations on the following 7 variables.

centre the study center, a factor with levels 1 and 2.

treatment the treatment arm, a factor with levels placebo and treatment.

gender a factor with levels female and male.

age the age of the patient.

status the respiratory status (response variable), a factor with levels poor and good.

month the month, each patient was examined at months 0, 1, 2, 3 and 4.

subject the patient ID, a factor with levels 1 to 111.

## Details

In each of two centres, eligible patients were randomly assigned to active treatment or placebo. During the treatment, the respiratory status (categorised poor or good) was determined at each of four, monthly visits. The trial recruited 111 participants (54 in the active group, 57 in the placebo group) and there were no missing data for either the responses or the covariates. The question of interest is to assess whether the treatment is effective and to estimate its effect.

Note that the data are in long form, i.e, repeated measurements are stored as additional rows in the data frame.

## Source

C. S. Davis (1991), Semi-parametric and non-parametric methods for the analysis of repeated measurements with applications to clinical trials. *Statistics in Medicine*, **10**, 1959–1980.

## Examples

```
data("respiratory", package = "HSAUR2")
mosaicplot(xtabs(~ treatment + month + status, data = respiratory))
```

# SEMI-PARAMETRIC AND NON-PARAMETRIC METHODS FOR THE ANALYSIS OF REPEATED MEASUREMENTS WITH APPLICATIONS TO CLINICAL TRIALS

CHARLES S. DAVIS

*Department of Preventive Medicine, University of Iowa, 2837 Steindler Building, Iowa City, IA 52242, U.S.A.*

## SUMMARY

Techniques applicable for the analysis of longitudinal data when the response variable is non-normal are not nearly as comprehensive as for normally-distributed outcomes. However, there have been several recent developments. Semi-parametric and non-parametric methodology for the analysis of repeated measurements is reviewed. The commonly encountered design in which, for each subject, one assesses a univariate response variable at multiple fixed time points, is considered. The types of outcomes considered include binary, ordered categorical, and continuous (but extremely non-normal) response variables. All of the methods considered allow for incomplete data due to the occurrence of missing observations. In addition, discrete and/or continuous covariates, which may be time-dependent, are accommodated by some of the approaches. The methods are demonstrated using data from three clinical trials.

## 1. INTRODUCTION

Consider a study comparing the effectiveness of two or more treatments, for example, a clinical trial that compares one or more experimental therapies with the standard treatment for a specific disease or condition. Such clinical trials often entail study designs that involve repeated observations on the same experimental unit. In this paper, I consider the commonly encountered design in which, for each subject, one assesses a univariate response variable at multiple points in time.

The analysis of data from such longitudinal clinical trials poses two main difficulties. First, the analysis is complicated by the dependence among successive observations made on the same individual. Second, since the investigators cannot usually control completely the circumstances for obtaining measurements, there may be incomplete data from individual subjects due to missing observations.

General approaches for the analysis of repeated measures are available for both continuous and categorical response variables. The basis for the classical methods of analysis pertains to continuous response variables and consists of parametric models that assume a multivariate normal error structure; these methods for normally-distributed responses appear in various review articles.<sup>1–3</sup> Although missing data and unbalanced patterns of observations may invalidate standard parametric analyses, there are available parametric methods to deal with missing data.<sup>4–7</sup> Koch *et al.*<sup>8</sup> were the first to develop a general approach to the analysis of repeated measures when the response is categorical and based their approach on the weighted least squares (WLS) methodology of Grizzle *et al.*<sup>9</sup> Various authors have extended this methodology to a



variety of response functions for complete and incomplete repeated measures categorical data.<sup>10-13</sup>

The above normal theory and categorical methods for repeated measures, however, do not always apply. First, the parametric assumptions underlying the classical analysis methods are frequently not tenable. In some studies, the response is continuous, but the distribution of the outcome variable is extremely non-normal. In addition, in situations in which the response is dichotomous or an ordered categorical variable, the general-purpose categorical methods often have limited usefulness. The WLS methodology allows for categorical covariates only, thus one cannot use it with continuous independent variables. In addition, it requires sufficient sample size for the marginal response functions at each time point within each category of the multi-way cross-classification of the covariates to have an approximately multivariate normal distribution. In practice, this imposes limitations on the total number of measurement times, the total number of covariates and the number of distinct levels of each covariate.

Several extensions of generalized linear models<sup>14</sup> to the analysis of repeated measures data have appeared.<sup>15-19</sup> These semi-parametric approaches are useful in longitudinal data analyses with univariate outcomes for which the quasi-likelihood formulation is sensible, for example, normal, Poisson, binomial and gamma response variables. The methods allow for missing observations and continuous (possibly time-dependent) covariates. Although the semi-parametric approaches are quite flexible, they still require assumptions concerning the distribution of the response variable. Thus, they may not apply to studies with a continuous, but extremely non-normal, response. Such situations may indicate the use of non-parametric methods. In addition to the substantial literature regarding distribution-free methods for complete multivariate observations,<sup>20-23</sup> non-parametric methods for incomplete repeated measurements are also available.<sup>24-28</sup> Two of these methods<sup>25,26</sup> have general applicability and, although developed for the special case of comparing two treatment groups, they do offer the advantage of allowing for differential patterns of missing observations in the two groups.

The purpose of this paper is to review and compare recent semi-parametric and non-parametric methods for analysis of repeated measurements. Section 2 describes three examples from clinical trials with repeated measures. The respective response variables are binary, ordered categorical and continuous (but extremely non-normal). Section 3 reviews and compares recent semi-parametric and non-parametric statistical methodology for analysis of repeated measurements. Section 4 considers application of the various methods to the three data sets. Section 5 concludes with recommendations, issues related to software availability and areas for further research.

## 2. EXAMPLES

### 2.1. Binary response

Appendix I displays the raw data from a clinical trial comparing two treatments for a respiratory illness.<sup>29</sup> In each of two centres, eligible patients were randomly assigned to active treatment or placebo. During treatment, respiratory status (categorized here as 0 = poor, 1 = good) was determined at four visits. Potential covariates were centre, sex and baseline respiratory status (all dichotomous), as well as age (in years) at the time of study entry. There were 111 patients (54 active, 57 placebo) with no missing data for responses or covariates.

### 2.2. Ordered categorical response

In a comparison of the effects of varying dosages of an anaesthetic on post-surgical recovery, 60 young children undergoing outpatient surgery were randomized to one of four dosages (15, 20,

Finally, even in the normal-theory parametric setting, lack of software for longitudinal data analysis is a major shortcoming. This problem is even more critical for the methods discussed here. Before we can recommend widespread usage of these non-parametric and semi-parametric approaches, we need to have user-friendly integrated software or, at the least, stand-alone programs.

## APPENDIX I: RESPIRATORY DISORDER EXAMPLE

Patient	Treatment	Sex	Age	Respiratory status (0 = poor, 1 = good)				
				Baseline	Visit 1	Visit 2	Visit 3	Visit 4
Centre 1								
1	P	M	46	0	0	0	0	0
2	P	M	28	0	0	0	0	0
3	A	M	23	1	1	1	1	1
4	P	M	44	1	1	1	1	0
5	P	F	13	1	1	1	1	1
6	A	M	34	0	0	0	0	0
7	P	M	43	0	1	0	1	1
8	A	M	28	0	0	0	0	0
9	A	M	31	1	1	1	1	1
10	P	M	37	1	0	1	1	0
11	A	M	30	1	1	1	1	1
12	A	M	14	0	1	1	1	0
13	P	M	23	1	1	0	0	0
14	P	M	30	0	0	0	0	0
15	P	M	20	1	1	1	1	1
16	A	M	22	0	0	0	0	1
17	P	M	25	0	0	0	0	0
18	A	F	47	0	0	1	1	1
19	P	F	31	0	0	0	0	0
20	A	M	20	1	1	0	1	0
21	A	M	26	0	1	0	1	0
22	A	M	46	1	1	1	1	1
23	A	M	32	1	1	1	1	1
24	A	M	48	0	1	0	0	0
25	P	F	35	0	0	0	0	0
26	A	M	26	0	0	0	0	0
27	P	M	23	1	1	0	1	1
28	P	F	36	0	1	1	0	0
29	P	M	19	0	1	1	0	0
30	A	M	28	0	0	0	0	0
31	P	M	37	0	0	0	0	0
32	A	M	23	0	1	1	1	1
33	A	M	30	1	1	1	1	0
34	P	M	15	0	0	1	1	0
35	A	M	26	0	0	0	1	0
36	P	F	45	0	0	0	0	0
37	A	M	31	0	0	1	0	0
38	A	M	50	0	0	0	0	0
39	P	M	28	0	0	0	0	0
40	P	M	26	0	0	0	0	0
41	P	M	14	0	0	0	0	1

# Analysing Longitudinal Data II – Generalised Estimation Equations and Linear Mixed Effect Models: Treating Respiratory Illness and Epileptic Seizures

## 13.1 Introduction

The data in Table 13.1 were collected in a clinical trial comparing two treatments for a respiratory illness (Davis, 1991).

**Table 13.1:** respiratory data. Randomised clinical trial data from patients suffering from respiratory illness. Only the data of the first seven patients are shown here.

centre	treatment	gender	age	status	month	subject
1	placebo	female	46	poor	0	1
1	placebo	female	46	poor	1	1
1	placebo	female	46	poor	2	1
1	placebo	female	46	poor	3	1
1	placebo	female	46	poor	4	1
1	placebo	female	28	poor	0	2
1	placebo	female	28	poor	1	2
1	placebo	female	28	poor	2	2
1	placebo	female	28	poor	3	2
1	placebo	female	28	poor	4	2
1	treatment	female	23	good	0	3
1	treatment	female	23	good	1	3
1	treatment	female	23	good	2	3
1	treatment	female	23	good	3	3
1	treatment	female	23	good	4	3
1	placebo	female	44	good	0	4
1	placebo	female	44	good	1	4
1	placebo	female	44	good	2	4
1	placebo	female	44	good	3	4
1	placebo	female	44	poor	4	4
1	placebo	male	13	good	0	5

**Table 13.1:** respiratory data (continued).

centre	treatment	gender	age	status	month	subject
1	placebo	male	13	good	1	5
1	placebo	male	13	good	2	5
1	placebo	male	13	good	3	5
1	placebo	male	13	good	4	5
1	treatment	female	34	poor	0	6
1	treatment	female	34	poor	1	6
1	treatment	female	34	poor	2	6
1	treatment	female	34	poor	3	6
1	treatment	female	34	poor	4	6
1	placebo	female	43	poor	0	7
1	placebo	female	43	good	1	7
1	placebo	female	43	poor	2	7
1	placebo	female	43	good	3	7
1	placebo	female	43	good	4	7
:	:	:	:	:	:	:

In each of two centres, eligible patients were randomly assigned to active treatment or placebo. During the treatment, the respiratory status (categorised poor or good) was determined at each of four, monthly visits. The trial recruited 111 participants (54 in the active group, 57 in the placebo group) and there were no missing data for either the responses or the covariates. The question of interest is to assess whether the treatment is effective and to estimate its effect.

form introduced in Chapter 12. For the **respiratory** data in Table 13.1 we could then apply logistic regression and for **epilepsy** in Table 13.2, Poisson regression. It can be shown that this approach will give *consistent* estimates of the regression coefficients, i.e., with large samples these point estimates should be close to the true population values. But the assumption of the independence of the repeated measurements will lead to estimated standard errors that are too small for the between-subjects covariates (at least when the correlation between the repeated measurements are positive) as a result of assuming that there are more independent data points than are justified.

We might begin by asking if there is something relatively simple that can be done to ‘fix-up’ these standard errors so that we can still apply the `R glm` function to get reasonably satisfactory results on longitudinal data with a non-normal response? Two approaches which can often help to get more suitable estimates of the required standard errors are *bootstrapping* and use of the *robust/sandwich*, *Huber-White variance estimator*.

The idea underlying the bootstrap (see Chapter 8 and Chapter 9), a technique described in detail in Efron and Tibshirani (1993), is to resample from the observed data with replacement to achieve a sample of the same size each time, and to use the variation in the estimated parameters across the set of bootstrap samples in order to get a value for the sampling variability of the estimate (see Chapter 8 also). With correlated data, the bootstrap sample needs to be drawn with replacement from the set of independent subjects, so that intra-subject correlation is preserved in the bootstrap samples. We shall not consider this approach any further here.

The sandwich or robust estimate of variance (see Everitt and Pickles, 2000, for complete details including an explicit definition), involves, unlike the bootstrap which is computationally intensive, a closed-form calculation, based on an asymptotic (large-sample) approximation; it is known to provide good results in many situations. We shall illustrate its use in later examples.

But perhaps more satisfactory would be an approach that fully utilises information on the data’s structure, including dependencies over time. In the linear mixed models for Gaussian responses described in Chapter 12, estimation of the regression parameters linking explanatory variables to the response variable and their standard errors needed to take account of the correlational structure of the data, but their interpretation could be undertaken independent of this structure. When modelling non-normal responses this independence of estimation and interpretation no longer holds. Different assumptions about how the correlations are generated can lead to regression coefficients with different interpretations. The essential difference is between *marginal models* and *conditional models*.

### 13.2.1 Marginal Models

Longitudinal data can be considered as a series of cross-sections, and marginal models for such data use the generalised linear model (see Chapter 7) to fit

## Respiratory Disorder (dichotomous)

STAT 22.  
Week 5

```
> library(HSAUR2) > data(respiratory)
> head(respiratory) #long form
  centre treatment gender age  status month subject
1       1 placebo female  46   poor     0         1
112      1 placebo female  46   poor     1         1
223      1 placebo female  46   poor     2         1
334      1 placebo female  46   poor     3         1
445      1 placebo female  46   poor     4         1
2       1 placebo female  28   poor     0         2

> table(gender, treatment)
      treatment
gender placebo treatment
female    200      240
male       85       30

> table(status, treatment, centre)
, , centre = 1
  treatment
status placebo treatment
poor          93         67
good          52         68

, , centre = 2
  treatment
status placebo treatment
poor          65         31
good          75        104
```

not 5 obs / person  
tub 1/4 & al  
56 cent 1, 55 cent 2

```
> #Data manip from HSAUR #The baseline status, i.e., the status for month == 0, needs to
  enter the models as an explanatory variable (HSAUR)
> #rearrange the data.frame respiratory in order to create a new variable baseline
> resp <- subset(respiratory, month > "0")
> resp$baseline <- rep(subset(respiratory, month == "0")$status, rep(4, 111))
> resp$nstat <- as.numeric(resp$status == "good")
> #new variable nstat is simply a dummy coding for a poor respiratory status
> resp$month <- resp$month[, drop = TRUE]
```

```
# ignore individual trajectories, compare mean outcomes across groups
> resp_glm <- glm(status ~ centre + treatment + gender + baseline + age, data = resp,
  family = "binomial")
> summary(resp_glm) # matches HSAUR
Coefficients:
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.900171	0.337653	-2.666	0.00768 **
centre2	0.671601	0.239567	2.803	0.00506 **
treatmenttreatment	1.299216	0.236841	5.486	4.12e-08 ***
gendermale	0.119244	0.294671	0.405	0.68572
baselinegood	1.882029	0.241290	7.800	6.20e-15 ***
age	-0.018166	0.008864	-2.049	0.04043 *

or use  
nstat

```
> exp(1.299) # odds 3.7 times higher of "good" for treat, not in love with adjusting initial status
[1] 3.665629
```

```
# now keep track of within-subject data with lmer (HSAUR does various gee also)
> resp_lmer <- lmer(status ~ baseline + month + treatment + gender + age + centre
  + (1 | subject), family = binomial(), data = resp)
> summary(resp_lmer) # allowing individ mean levels to differ, no trend apparent
Generalized linear mixed model fit by the Laplace approximation
Formula: status ~ baseline + month + treatment + gender + age + centre + (1 | subject)
```

Random effects:

Groups	Name	Variance	Std.Dev.
subject	(Intercept)	3.9739	1.9935

Number of obs: 444, groups: subject, 111  
Fixed effects:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.66664	0.76708	-2.173	0.0298 *
baselinegood	3.10734	0.53246	5.836	5.35e-09 ***
month.L	-0.22795	0.27186	-0.838	0.4018
month.Q	-0.03890	0.27158	-0.143	0.8861
month.C	-0.36892	0.27269	-1.353	0.1761
treatmenttreatment	2.18393	0.52365	4.171	3.04e-05 ***
gendermale	0.20448	0.66883	0.306	0.7598
age	-0.02566	0.02021	-1.269	0.2043
centre2	1.05614	0.53808	1.963	0.0497 *

```
> exp(fixef(resp_lmer))
(Intercept) baselinegood month.L month.Q month.C treatmenttreatment
0.1888801 22.3614175 0.7961675 0.9618516 0.6914812 8.8811669
gendermale age centre2
1.2268855 0.9746698 2.8752589
```

use nstat  
same resultodds (good)  
9x larger  
in treat

```

> library(HSAUR2) > data(respiratory)
> head(respiratory) #long form
  centre treatment gender age  status month subject
1       1 placebo female 46   poor     0        1
112      1 placebo female 46   poor     1        1
223      1 placebo female 46   poor     2        1
334      1 placebo female 46   poor     3        1
445      1 placebo female 46   poor     4        1
2       1 placebo female 28   poor     0        2
> table(status, treatment, centre)
, , centre = 1
   treatment
status placebo treatment
poor         93         67
good         52         68
, , centre = 2
   treatment
status placebo treatment
poor         65         31
good         75        104

> #Data manip from HSAUR #The baseline status, i.e., the status for month == 0, needs to
  enter the models as an explanatory variable (HSAUR)
> #rearrange the data.frame respiratory in order to create a new variable baseline
> resp <- subset(respiratory, month > "0")
> resp$baseline <- rep(subset(respiratory, month == "0")$status, rep(4, 111))
> resp$nstat <- as.numeric(resp$status == "good")
> #new variable nstat is simply a dummy coding for a poor respiratory status
> resp$month <- resp$month[, drop = TRUE]

# ignore individual trajectories, compare mean outcomes across groups
> resp_glm <- glm(status ~ centre + treatment + gender + baseline + age, data = resp,
  family = "binomial")
> summary(resp_glm) # matches HSAUR
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.900171   0.337653  -2.666   0.00768 **
centre2         0.671601   0.239567   2.803   0.00506 **
treatmenttreatment 1.299216   0.236841   5.486 4.12e-08 ***
gendermale      0.119244   0.294671   0.405   0.68572
baselinegood    1.882029   0.241290   7.800 6.20e-15 ***
age            -0.018166   0.008864  -2.049   0.04043 *
---
> exp(1.299)# odds 3,7 times higher of "good" for treat, not in love with adjusting initial status
[1] 3.665629

# now keep track of within-subject data with lmer (HSAUR does various gee also)
> resp_lmer <- lmer(status ~ baseline + month + treatment + gender + age + centre
  + (1 | subject), family = binomial(), data = resp)
> summary(resp_lmer) # allowing individ mean levels to differ, no trend apparent
Generalized linear mixed model fit by the Laplace approximation
Formula: status ~ baseline + month + treatment + gender + age + centre + (1 | subject)
Random effects:
 Groups Name Variance Std.Dev.
subject (Intercept) 3.9739 1.9935
Number of obs: 444, groups: subject, 111
Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.66664   0.76708  -2.173   0.0298 *
baselinegood    3.10734   0.53246   5.836 5.35e-09 ***
month.L        -0.22795   0.27186  -0.838   0.4018
month.Q        -0.03890   0.27158  -0.143   0.8861
month.C        -0.36892   0.27269  -1.353   0.1761
treatmenttreatment 2.18393   0.52365   4.171 3.04e-05 ***
gendermale      0.20448   0.66883   0.306   0.7598
age            -0.02566   0.02021  -1.269   0.2043
centre2         1.05614   0.53808   1.963   0.0497 *
---
> exp(fixef(resp_lmer))
  (Intercept) baselinegood month.L month.Q month.C treatmenttreatment
1.1888801    22.3614175    0.7961675 0.9618516 0.6914812      8.8811669
1.2268855     0.9746698    2.8752589
Exponentiate endpoints of confint

```



```
# we did this format in class ex (from 2013); need to change syntax slightly in 2014
> resp_lmera = lmer(status ~ treatment + (1| subject), family = binomial(), data = resp)
Warning message:
In lmer(status ~ treatment + (1 | subject), family = binomial(), :
  calling lmer with 'family' is deprecated; please use glmer() instead

# 2014 version of syntax
# In terms of our modeling: Level 1 is simply the mean (no trend) param alph_0. Level 2 says alph_0 dep
> resp_lmera = glmer(status ~ treatment + (1| subject), family = binomial, data = resp)
> summary(resp_lmera)
Generalized linear mixed model fit by maximum likelihood ['glmerMod']
Family: binomial ( logit )
Formula: status ~ treatment + (1 | subject)
Data: resp

      AIC      BIC    logLik deviance
482.6303 494.9178 -238.3152 476.6303

Random effects:
Groups Name      Variance Std.Dev.
subject (Intercept) 6.4      2.53
Number of obs: 444, groups: subject, 111

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.4714    0.3875  -1.217 0.223775
treatmenttreatment  2.0533    0.5660   3.628 0.000286 *** # only slightly larger standard error than
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
              (Intr)
trtmnttrtmn -0.685

> exp(fixef(resp_lmera))
      (Intercept) treatmenttreatment
      0.624135      7.793621 # not far from the 8.8 from the class ex

> confint(resp_lmera, method = "boot", nsim = 1000, boot.type = "perc")
Computing bootstrap confidence intervals ...
              2.5 %      97.5 %
sd_(Intercept)|subject 1.7460690 3.2564314
(Intercept)          -1.3960714 0.3816819
treatmenttreatment    0.9364879 3.6391380
> exp(.9365)
[1] 2.551037 # endpoints of 95% CI for increase in odds of 'good' going from placebo to control
> exp(3.639)
[1] 38.05376

> # gender issues
# One way to look at this is to make the Level 2 model a 2x2 factorial design, where the genderXtreame
the differential effectiveness for males and females

> resp_lmeraG = glmer(status ~ gender*treatment + (1| subject), family = binomial, data = resp)
```



# Power for linear models of longitudinal data with applications to Alzheimer's Disease Phase II study design

Michael C. Donohue, Steven D. Edland, Anthony C. Gamst  
Division of Biostatistics and Bioinformatics  
University of California, San Diego

January 22, 2013

## 1 Introduction

We will discuss power and sample size estimation for randomized placebo controlled studies in which the primary inference is based on the interaction of treatment and time in a linear mixed effects model (Laird and Ware 1982). We will demonstrate how the sample size formulas of Liu and Liang (1997) for marginal or generalized estimating equation (GEE) models (Zeger and Liang 1986) can be adapted for mixed effects models. Finally, using mixed effects model estimates based on data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), we will give examples of sample size calculations for models with and without baseline covariates which may help explain heterogeneity in cognitive decline and improve power.

## 2 Power calculations

### 2.1 Exchangeable correlation and random intercept models

Suppose we wish to estimate the required sample size for inference regarding the interaction of treatment and time in a longitudinal, placebo controlled

and  $d = 0.5$  as follows:

```

n = 3
t = c(0, 2, 5)
u = list(u1 = t, u2 = rep(0, n))
v = list(v1 = cbind(1, 1, rep(0, n)),
        v2 = cbind(1, 0, t))
rho = c(0.2, 0.5, 0.8)
sigma2 = c(100, 200, 300)
tab = outer(rho, sigma2,
            Vectorize(function(rho, sigma2){
              round(diggle.linear.power(
                d=0.5,
                t=t,
                sigma2=sigma2,
                R=rho,
                alternative="one.sided",
                power=0.80)$n)))
colnames(tab) = paste("sigma2 =", sigma2)
rownames(tab) = paste("rho =", rho)
tab

```

	sigma2 = 100	sigma2 = 200	sigma2 = 300
rho = 0.2	312	625	937
rho = 0.5	195	390	586
rho = 0.8	78	156	234

As a second example, consider an Alzheimer's disease trial in which assessments are taken every three months for 18 months (7 visits). We assume an smallest detectable effect size of 1.5 points on the cognitive portion of the Alzheimer's Disease Assessment Scale (ADAS-Cog). This is a 70 point scale with great variability among sick individuals. We assume the random intercept to have a variance of 55, the random slope to have a variance of 24, and a residual variance of 10. The correlation between random slope term and random intercept term is 0.8. We can estimate the necessary sample size by first generating the correlation structure. Since  $\varepsilon = \text{var}(Y_{ij})$  is not constant over time in this model, we fix  $\text{sigma2}=1$  and set  $R$  equal to the covariance matrix for  $\varepsilon_i$ :

```

# var of random intercept
sig2.i = 55
# var of random slope
sig2.s = 24
# residual var
sig2.e = 10

```

see RQ1

```
# covariance of slope and intercep
cov.s.i <- 0.8*sqrt(sig2.i)*sqrt(sig2.s)
cov.t <- function(t1, t2, sig2.i, sig2.s, cov.s.i){
  sig2.i + t1*t2*sig2.s + (t1+t2)*cov.s.i
}
t = seq(0,1.5,0.25)
n = length(t)
R = outer(t, t, function(x,y){cov.t(x,y, sig2.i, sig2.s, cov.s.i)})
R = R + diag(sig2.e, n, n)
u = list(u1 = t, u2 = rep(0,n))
v = list(v1 = cbind(1,1,rep(0,n)),
        v2 = cbind(1,0,t))
liu.liang.linear.power(d=1.5, u=u, v=v, R=R, sig.level=0.05, alternative="two.sided")
```

Longitudinal linear model power calculation (Liu & Liang, 1997)

```
N = 414.6202
n = 207.3101, 207.3101
delta = 1.5
sigma2 = 1
sig.level = 0.05
power = 0.8
alternative = two.sided
```

NOTE: N is total sample size and n is sample size in each group.

R:

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]
[1,]	65.00000	62.26636	69.53272	76.79908	84.06544	91.3318	98.59817
[2,]	62.26636	81.03272	79.79908	88.56544	97.33180	106.0982	114.86453
[3,]	69.53272	79.79908	100.06544	100.33180	110.59817	120.8645	131.13089
[4,]	76.79908	88.56544	100.33180	122.09817	123.86453	135.6309	147.39725
[5,]	84.06544	97.33180	110.59817	123.86453	147.13089	150.3972	163.66361
[6,]	91.33180	106.09817	120.86453	135.63089	150.39725	175.1636	179.92997
[7,]	98.59817	114.86453	131.13089	147.39725	163.66361	179.9300	206.19633

So the study would require about 207 subjects per arm to achieve 80% power, with a two-tailed  $\alpha = 0.05$ .

The simple formula provided in Diggle et al. 2002 suggests the required number of subjects can be found by  $2(z_\alpha + 2Q)\xi/d^2$ , where

$$\xi_{\text{WRONG}} = \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & \cdots & 1 \\ t_1 & \cdots & t_n \end{pmatrix} R^{-1} \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_n \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

# Sample Size Planning for Longitudinal Models: Accuracy in Parameter Estimation for Polynomial Change Parameters

Ken Kelley

University of Notre Dame

Joseph R. Rausch

Cincinnati Children's Hospital Medical Center and University of  
Cincinnati College of Medicine

Longitudinal studies are necessary to examine individual change over time, with group status often being an important variable in explaining some individual differences in change. Although sample size planning for longitudinal studies has focused on statistical power, recent calls for effect sizes and their corresponding confidence intervals underscore the importance of obtaining sufficiently accurate estimates of group differences in change. We derived expressions that allow researchers to plan sample size to achieve the desired confidence interval width for group differences in change for orthogonal polynomial change parameters. The approaches developed provide the expected confidence interval width to be sufficiently narrow, with an extension that allows some specified degree of assurance (e.g., 99%) that the confidence interval will be sufficiently narrow. We make computer routines freely available, so that the methods developed can be used by researchers immediately.

*Keywords:* sample size planning, research design, accuracy in parameter estimation, longitudinal data analysis, group comparisons

Longitudinal studies have become a major source of knowledge generation in psychology and related disciplines. This is the case in part because of the rich information inherently provided by repeated measurement of the same set of individuals over time, as well as the sophisticated methods developed over the last three decades that allow a wide variety of questions about intraindividual change and interindividual differences in change to be addressed (see, for example, Collins & Horn, 1991; Collins & Sayer, 2001; Fitzmaurice, Davidian, Verbeke, & Molenberghs, 2009; Singer & Willett, 2003, for reviews of longitudinal data analytic methods). While the analysis of longitudinal data gained widespread usage in psychology and related disciplines, comparisons of mean differences across groups continue to be widely used. Naturally, the idea of examining group differences over time itself became a widely used technique. Examining group-by-time interactions allows researchers to infer (a) whether groups are changing differently and (b) by how much groups are changing differently.

The question of “are groups changing differently” functionally is answered in a dichotomous manner via the results of a null hypothesis significance test. Namely, if the  $p$  value is less than the

specified Type I error rate (e.g., .05), the null hypothesis of groups changing the same over time (i.e., the group-by-time interaction) is rejected, with the conclusion being that groups do indeed change differently. However, if the  $p$  value is greater than the specified Type I error rate, the null hypothesis is not rejected. Of course, the failure to reject a null hypothesis does not imply that the null hypothesis is in fact true. However, in such cases, the failure to find statistical significance at least does not show support for a difference. Obtaining a clear answer to the research question “are groups changing differently” is functionally answered when the null hypothesis of the group-by-time interaction is rejected.

The question of “by how much do groups change differently” is not answered with a null hypothesis significance test, but rather it is addressed continuously on the basis of a point estimate of the group-by-time interaction and the corresponding confidence interval for the population value. The magnitude of the group-by-time interaction, that is, how different the slopes of two groups are, is often an important outcome in longitudinal studies. Additionally, there is a one-to-one relationship between two-sided  $(1 - \alpha)100\%$  confidence interval and a nondirectional null hypothesis significance test with a Type I error rate of  $\alpha$ .<sup>1</sup> Namely, if the value of the specified null hypothesis (e.g., 0) is not contained within the  $(1 - \alpha)100\%$  confidence interval limits, that same value would be rejected as the value of the null hypothesis using a Type I error rate of  $\alpha 100\%$ . Thus, it is known that a particular null hypothesis will be rejected if the corresponding confidence interval does not contain the specified null value. However, because the confidence interval contains those values that cannot be rejected as implausi-

---

Ken Kelley, Department of Management, Mendoza College of Business, University of Notre Dame; Joseph R. Rausch, Division of Behavioral Medicine and Clinical Psychology, Cincinnati Children's Hospital Medical Center and University of Cincinnati College of Medicine.

This research was supported in part by the University of Notre Dame Center for Research Computing through the use of the high performance computing cluster.

Correspondence concerning this article should be addressed to Ken Kelley, Department of Management, Mendoza College of Business, University of Notre Dame, Notre Dame, IN 46556. E-mail: kkelley@nd.edu

---

<sup>1</sup> There is also an analogous relationship between a one-sided confidence interval and a directional hypothesis tests. As noted, if a confidence interval contains the specified null value, then the corresponding null hypothesis, with the same value, will be rejected.

limited by the appropriateness of the value(s) used for the population parameter(s). The appropriateness of the sample size planning procedure output to the misspecification of the population parameter input speaks to the robustness of the procedure. Although the robustness of a statistical procedure is important, it is beyond the scope of this article. What is important here is that the proposed sample size planning methods perform as they should, given that the appropriate values for the population parameters have been specified. That is, under the ideal circumstances, we seek to answer the question “Does the method we propose perform in an optimal way?” We used our Monte Carlo simulation study to evaluate the appropriateness of the procedure given that the correct values are supplied. We now outline the two studies we used for the bases of the parameter values.

### **Study 1: Tolerance of Antisocial Thinking During Adolescence**

We used the Elliot, Huizinga, and Menard (1989) study of the tolerance of antisocial thinking during adolescence as the basis for part of our Monte Carlo simulation study. The Elliot et al. (1989) used data from the National Youth Survey, where a dependent variable of interest was “tolerance of antisocial behavior” and five measurement occasions for Cohort 1 (age 11 years at the beginning and 15 years at the end of the study) with a sample size of 239. This study was also used by Raudenbush and Xiao-Feng (2001) in the context of power analysis, as well as in other works to illustrate various methodological issues (e.g., Miyazaki & Raudenbush, 2000; Raudenbush & Chan, 1992, 1993; Willett & Sayer, 1994). Like Raudenbush and Xiao-Feng (2001), we used sex as a grouping variable so as to estimate the group-by-time (i.e., sex-by-time) interaction (i.e.,  $\beta_{11}$ ). The estimates used are those reported in Raudenbush and Xiao-Feng (2001): the error variance (i.e.,  $\hat{\sigma}^2$ ) is 0.0262, the true variance of the intercept is 0.0333, and the true variance of the slope is .0030. Note that it is not necessary to specify a population value or a parameter of minimal interest for the slope, as is the case in power analysis, as the width of the confidence interval is independent of the value of the slope.<sup>10</sup>

For the tolerance of antisocial thinking during adolescence data, a 3 (number of measurement occasions = 3, 5, and 10) by 2 (widths = .25 and .05) by 2 (sample size procedure for the expected confidence interval width will be sufficiently narrow and there will be 85% assurance that the confidence interval will be sufficiently narrow) Monte Carlo simulation was used. Our reading of the literature suggested that there often tend to be fewer rather than many measurement occasions in psychology and related disciplines. The number of measurement occasions of three, five, and 10 seemed quite reasonable, given what has been found in the literature for typical longitudinal designs (e.g., Kwok, West, & Green, 2007).

Each of the 12 conditions was based on 10,000 replications using the PROC MIXED procedure in SAS (Version 9.2). Such a large number of replications were used so that we could very accurately estimate the mean and median confidence interval widths for the expected width case and the percentile and percentile rank of the desired width for the assurance case. In the Monte Carlo simulation, all assumptions were satisfied, illustrating the ideal conditions in order to evaluate the effectiveness of our sample size planning methods. Our combination of conditions led

to planned sample sizes that ranged from small (e.g., 43) to relatively large (e.g., 793); demonstrating the relative variety of situations of the conditions used in the Monte Carlo simulation study to examine the effectiveness of the sample size planning procedures.

Table 1 shows the results of the Monte Carlo simulation based on the tolerance of antisocial thinking during adolescence data for the expected confidence interval width. As Table 1 shows, the mean and median confidence interval widths were nearly identical to the desired width in most cases. The biggest discrepancy was for the widest confidence interval condition, where the desired width ( $\omega$ ) was 0.05 and necessary sample size was only 43 per group. In this most discrepant condition, the mean of the confidence interval widths was 0.0487, illustrating the mean confidence interval widths that were 0.0013 units smaller than specified. As the sample sizes became larger, the desired width and the empirical widths converged and became nearly identical. Thus, in this situation, the procedure developed for planning sample size to ensure that the expected width would be sufficiently narrow worked very well.

Table 2 shows the results of the Monte Carlo simulation based on the tolerance of antisocial thinking during adolescence data when an assurance parameter is incorporated produced the desired proportion of confidence intervals that were sufficiently narrow no less than the specified assurance of .85. The biggest discrepancy was again for the 0.05 condition, where the procedure implied sample size was 49. Analogous to the expected width situation, as the sample size becomes larger, the empirical assurance approaches the specified value. Thus, in this situation, the procedure developed for planning sample size to provide a desired degree of assurance worked very well.

### **Study 2: Quality of Marriage**

Karney and Bradbury (1995) provided a tutorial on how change models can be used to better understand the way in which the quality of marriage changes over time that is based on repeatedly measuring the same set of married individuals. Karney and Bradbury (1995) provided illustrative data from a study of newlywed couples. In particular, the data were from 25 newlywed wives from five measurement occasions over the first 30 months of marriage (measured approximately every 6 months), where the participants self-reported marital quality using the Marital Adjustment Test (MAT; Locke & Wallace, 1959). In general, a sample size of 25 is inordinately small for an application of a multilevel change model. However, we used their data simply for illustrative purposes, where the estimate of the error variance (i.e.,  $\hat{\sigma}_e^2$ ) is 134.487, the estimate of the true variance of the intercept is 447.393, and the

<sup>10</sup> For power analysis, a value for the group-by-time interaction, or a standardized version which implicitly includes the slope as well as the variance, must be specified, as the noncentral parameter depends on it. However, because the confidence interval width is independent of the variability, as is the case for a normal distribution, the slope is not specified in the AIPE approach to sample size planning. This is true for AIPE whenever the effect size is independent of its variance, which is not the case for all effect sizes (e.g., standardized mean difference, coefficient of variation, squared multiple correlation coefficient, and so on). Thus, because one less parameter value needs to be specified in the AIPE approach, it is easier to plan sample size from an AIPE perspective.

- Rogosa, D. R., & Saner, H. M. (1995). Longitudinal data analysis examples with random coefficient models. *Journal of Educational and Behavioral Statistics*, 20, 149–170.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581–592.
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.
- Stuart, A., Ord, J. K., & Arnold, S. (1999). *Kendall's advanced theory of statistics: Vol. 2A. Classical inference and the linear model* (6th ed.). New York, NY: Oxford University Press.
- Venables, W. N., Smith, D. M., & the R Development Core Team. (2010). *An introduction to R*. Vienna, Austria: The R Development Core Team.
- Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin*, 116, 363–381.

## Appendix

### Using R and MBESS to Implement the Methods Discussed

A function for sample size planning from the AIPE perspective for polynomial change models was written and incorporated into the MBESS (Kelley, 2007a, 2007b, 2007c; Kelley & Lai, 2010) R package (R Development Core Team, 2010).<sup>12</sup>

Throughout the appendix, sans serif font denotes R functions, options of functions, or output. Sans serif font followed by an open parenthesis and immediately by a closed parenthesis denotes a particular R function. When specifications are given within the parentheses of a function, that function is directly executable in R, after the MBESS packages have been installed and loaded. The easiest way to install MBESS is with the `install.packages()` function in the following manner

```
install.packages(pkgs="MBESS")
```

assuming that the machine has an active Internet connection, which may require the user to select one of many download (i.e., mirror) sites. Alternatively, MBESS can be installed via the Package Manager drop-down menu (in the Windows and Macintosh versions) from the R toolbar, where the user selects from the many packages available to install onto his or her system. After MBESS is installed, it is loaded into the current session with the `require()` function, which is implemented as follows:

```
require(MBESS).
```

A set of help files also accompanies MBESS. For any function in MBESS (or R more generally), the help file can be displayed with the `help` function, `help()`. For example, the associated help files for the `ss.aipe.pcm()` function, the function that implements the sample size planning methods developed in the article, `help(ss.aipe.pcm)`.

Additionally, when the exact name of a function is not known, one can search for functions and help files by using the `help.search()` function. For example, if one were interested in computing a covariance matrix on a data set, searching for "covariance matrix" via the `help.search()` function as follows

```
help.search("covariance matrix")
```

returns information on functions that pertain to covariance matrices. More details on the way in which R is installed and used is available

for download via the freely available book *An Introduction to R* (Venables, Smith, & the R Development Core Team, 2010).

For the `ss.aipe.pcm()` function, which is the function that implements the methods developed in this article, the parameters of the function are

```
ss.aipe.pcm(true.variance.trend, error.variance, variance.true.minus.estimated.trend=NULL, duration, frequency, width, conf.level=.95, trend="linear", assurance=NULL),
```

some of which need to be specified on any implementation of the function, where `true.variance.trend` is the variance of the individuals' true change coefficients (i.e.,  $\sigma_{v_m}^2$ , the first component on the right-hand side of Equation 18), `error.variance` is the true error variance (i.e.,  $\sigma_e^2$  from the numerator of the right-hand side of Equation 19), and `variance.true.minus.estimated.trend` is the variance of the difference between the  $m$ th true change coefficient minus the  $m$ th estimated change coefficient (i.e.,  $\sigma_{\pi_m - \pi_m}^2$  from Equation 19). Because of the one-to-one relationship between  $\sigma_e^2$  and  $\sigma_{\pi_m - \pi_m}^2$ , only one of the two values needs to be specified. Further, the parameters of duration, frequency, width, confidence level (e.g., .90, .95, .99, and so forth), trend (either linear, quadratic, or cubic), and assurance (e.g., NULL for only an expected width, .85, .95, .99, and so forth) each need to be specified.

To illustrate how the `ss.aipe.pcm()` MBESS function is used, we will use the previously discussed tolerance of antisocial behavior example from Elliot et al. (1989), which was used as an exemplar by Raudenbush and Xiao-Feng (2001) for their contribution on sample size planning in the context of polynomial change model for the power analytic approach.

<sup>12</sup> R and MBESS are both open source and thus freely available. R is available for download via the Comprehensive R Archival Network (CRAN; <http://www.r-project.org/>) for computers running Microsoft Windows, Linux/Unix, and Apple Macintosh operating systems. The direct link to the MBESS page on CRAN, where the most up-to-date version of MBESS and its corresponding manual are available, is <http://cran.r-project.org/web/packages/MBESS/index.html> (note that these Internet addresses are case sensitive).



Suppose that a researcher would like to plan sample size so that the straight-line change coefficient has an expected 95% confidence interval width of 0.025 units, which the researcher believes is sufficiently narrow for the purposes of establishing an accurate difference between a treatment group and a control group. The study will have a duration of 4 years with one measurement occasions per year, for a total of five measurement occasions. The supposed variance of the linear trend (i.e.,  $\sigma_{\nu_m}^2$ ) of 0.003 and the supposed error variance ( $\sigma_e^2$ ) of 0.0262, both of which are obtained from literature (i.e., in Raudenbush & Xiao-Feng, 2001 based on the data of Elliot et al., 1989).

In this situation, the way in which the `ss.aipe.pcm()` MBESS function is implemented, after MBESS has been installed and loaded via the `require()` function, is as follows

```
ss.aipe.pcm(true.variance.trend=0.003,
  error.variance=0.0262, duration=4,
  frequency=1, width=0.025, conf.level=.95),
```

which returns the following output

“Results for expected width to be sufficiently narrow”  
278.

Thus, a sample size of 278 is required when the duration of the study will be 4 units and the frequency of measurement occasions is 1 year in order for the expected confidence interval width to be 0.025 units.

Suppose that the researcher was not happy with having *only* an expected confidence interval width for the group-by-time interaction of 0.025 units. Rather, suppose that the researcher wanted to have 99% assurance that the 95% confidence interval would be sufficiently narrow. The way in which sample size can be planned

in this situation with the `ss.aipe.pcm()` MBESS function is as follows,

```
ss.aipe.pcm(true.variance.trend=.003,
  error.variance=.0262,
  duration=4, frequency=1, width=.025,
  conf.level=.95, assurance=.99),
```

which returns the following output  
“Results for Assurance”  
316.

Thus, a sample size of 316 will be required to ensure that the 95% confidence interval will be sufficiently narrow (i.e., have a width less than .025 units) at least 99% of the time.

As can be seen, the functions are easy to use and require only minimal knowledge of R. Even if R will not be used for the analysis of the results, R can easily be used for sample size planning purposes. An additional function in the MBESS R package is the `ss.power.pcm()` function, which implements sample size planning for statistical power in this context. That is, the `ss.power.pcm()` function implements the methods developed by Raudenbush and Xiao-Feng (2001) for planning sample size in order to have a desired statistical power. Detailed information on the `ss.power.pcm()` function is available in the MBESS manual or from R via the command

```
help(ss.power.pcm)
```

after MBESS has been installed and loaded.

Received February 5, 2010

Revision received January 19, 2011

Accepted January 26, 2011 ■

see RQ1



A Community Site for R – Sponsored by Revolution Analytics



## ss.power.pcm {MBESS}

### Sample size planning for power for polynomial change models

**Package:** MBESS

**Version:** 3.3.3

### Description

Returns power given the sample size, or sample size given the desired power, for polynomial change models

### Usage

```
ss.power.pcm(beta, tau, level . 1. variance, frequency, duration, desired.power = NULL, N = NU
```

### Arguments

#### beta

the level two regression coefficient for the group by time interaction; where "X" is coded -.5 and .5 for the two groups.

#### tau

the true variance of the individuals' slopes

#### level . 1. variance

level one variance

#### frequency

frequency of measurements per unit of time duration of the study in the particular units (e.g., age, hours, grade level, years, etc.)

#### duration

time in some number of units (e.g., years)

#### desired.power

desired power

#### N

sample size

#### alpha.level

Type I error rate

#### standardized

the standardized slope is the unstandardized slope divided by the square root of tau, the variance of the unique effects for beta.

#### directional



should a one (TRUE) or two (FALSE) tailed test be performed.

## References

Raudenbush, S. W., & X-F., Liu. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods*, 6, 387--401.

## Examples

```
# Example from Raudenbush and Liu (2001)
# ss.power.pcm(beta=-.4, tau=.003, level.1.variance=.0262, frequency=2, duration=2, desired
# ss.power.pcm(beta=-.4, tau=.003, level.1.variance=.0262, frequency=2, duration=2, N=238, .

# The standardized effect size is obtained as beta/sqrt(tau): -.4/sqrt(.003) = -.0219.
# ss.power.pcm(beta=-.0219, tau=.003, level.1.variance=.0262, frequency=2, duration=2, desi
# ss.power.pcm(beta=-.0219, tau=.003, level.1.variance=.0262, frequency=2, duration=2, N=23
```

## Author(s)

Ken Kelley (University of Notre Dame; [KKelley@ND.Edu](mailto:KKelley@ND.Edu))

Documentation reproduced from package **MBESS, version 3.3.3**. License: GPL ( $\geq 2$ )

---

also see vignette

## Package ‘powerlmm’

August 14, 2018

**Type** Package

**Title** Power Analysis for Longitudinal Multilevel Models

**Version** 0.4.0

**Description** Calculate power for the 'time x treatment' effect in two- and three-level multilevel longitudinal studies with missing data. Both the third-level factor (e.g. therapists, schools, or physicians), and the second-level factor (e.g. subjects), can be assigned random slopes. Studies with partially nested designs, unequal cluster sizes, unequal allocation to treatment arms, and different dropout patterns per treatment are supported. For all designs power can be calculated both analytically and via simulations. The analytical calculations extends the method described in Galbraith et al. (2002) <doi:10.1016/S0197-2456(02)00205-2>, to three-level models. Additionally, the simulation tools provides flexible ways to investigate bias, Type I errors and the consequences of model misspecification.

**License** GPL (>= 3)

**URL** <https://github.com/rpsychologist/powerlmm>

**BugReports** <https://github.com/rpsychologist/powerlmm/issues>

**Encoding** UTF-8

**LazyData** true

**RoxygenNote** 6.1.0

**Depends** R (>= 3.2.0)

**Imports** stats, methods, parallel, lme4 (>= 1.1), Matrix, MASS, scales, utils

**Suggests** testthat, dplyr, tidyr, knitr, rmarkdown, pbmcapply (>= 1.1), lmerTest (>= 2.0), ggplot2 (>= 2.2), ggsci, viridis, gridExtra, shiny (>= 1.0), shinydashboard

**ByteCompile** true

**VignetteBuilder** knitr

**NeedsCompilation** no

---

`get_power`*Calculate power for two- and three-level models with missing data.*

---

## Description

Calculate power for two- and three-level models with missing data.

## Usage

```
get_power(object, df = "between", alpha = 0.05, progress = TRUE,
          R = 1L, cores = 1L, ...)
```

## Arguments

<code>object</code>	An object created by <a href="#">study_parameters</a>
<code>df</code>	Either "between" or, "satterth" for Satterthwaite's DF approximation. Also accepts a numeric value which will be used as DF.
<code>alpha</code>	The alpha level, defaults to 0.05.
<code>progress</code>	logical; displays a progress bar when > 1 power analysis is performed.
<code>R</code>	An integer indicating how many realizations to base power on. Useful when dropout or cluster sizes are sampled (i.e. are random variables).
<code>cores</code>	An integer indicating how many CPU cores to use.
<code>...</code>	Other potential arguments; currently used to pass progress bar from Shiny

## Details

### Calculation of the standard errors

Designs with equal cluster sizes, and with no missing data, uses standard closed form equations to calculate standard errors. Designs with missing data or unequal cluster sizes uses more computationally intensive linear algebra solutions.

To see a more detailed explanation of the calculations, type `vignette("technical", package = "powerlmm")`.

### Degrees of freedom

Power is calculated using the  $t$  distribution with non-centrality parameter  $b/se$ , and  $dfs$  are either based on a the between-subjects or between-cluster  $dfs$ , or using Satterthwaite's approximation. For the "between" method,  $N_3 - 2$  is used for three-level models, and  $N_2 - 2$  for two-level models, where  $N_3$  and  $N_2$  is the total number of clusters and subjects in both arms.

**N.B** Satterthwaite's method will be RAM and CPU intensive for large sample sizes. The computation time will depend mostly on  $n1$  and  $n2$ . For instance, for a fully nested model with  $n1 = 10$ ,  $n2 = 100$ ,  $n3 = 4$ , computations will likely take 30-60 seconds.

### Cluster sizes or dropout pattern that are random (sampled)

If `deterministic_dropout = FALSE` the proportion that dropout at each time point will be sampled from a multinomial distribution. However, if it is `TRUE`, the proportion of subjects that

---

## Foreword

I'm delighted to see this new book on multiple imputation by Stef van Buuren for several reasons. First, to me at least, having another book devoted to multiple imputation marks the maturing of the topic after an admittedly somewhat shaky initiation. Stef is certainly correct when he states in Section 2.1.2: "The idea to create multiple versions must have seemed outrageous at that time [late 1970s]. Drawing imputations from a distribution, instead of estimating the 'best' value, was a severe breach with everything that had been done before." I remember how this idea of multiple imputation was even ridiculed by some more traditional statisticians, sometimes for just being "silly" and sometimes for being hopelessly inefficient with respect to storage demands and outrageously expensive with respect to computational requirements.

Some others of us foresaw what was happening to both (a) computational storage (I just acquired a 64 GB flash drive the size of a small finger for under \$60, whereas only a couple of decades ago I paid over \$2500 for a 120 KB hard-drive larger than a shoe box weighing about 10 kilos), and (b) computational speed and flexibility. To develop statistical methods for the future while being bound by computational limitations of the past was clearly inapposite. Multiple imputation's early survival was clearly due to the insight of a younger generation of statisticians, including many colleagues and former students, who realized future possibilities.

A second reason for my delight at the publication of this book is more personal and concerns the maturing of the author, Stef van Buuren. As he mentions, we first met through Jan van Rijkevorsel at TNO. Stef was a young and enthusiastic researcher there, who knew little about the kind of statistics that I felt was essential for making progress on the topic of dealing with missing data. But consider the progress over the decades starting with his earlier work on MICE! Stef has matured into an independent researcher making important and original contributions to the continued development of multiple imputation.

This book represents a "no nonsense" straightforward approach to the application of multiple imputation. I particularly like Stef's use of graphical displays, which are badly needed in practice to supplement the more theoretical discussions of the general validity of multiple imputation methods. As I have said elsewhere, and as implied by much of what is written by Stef, "It's not that multiple imputation is so good; it's really that other methods for addressing missing data are so bad." It's great to have Stef's book on mul-

# Package ‘mice’

March 25, 2012

**Type** Package

**Version** 2.12

**Title** Multivariate Imputation by Chained Equations

**Date** 2012-03-25

**Author** Stef van Buuren <stef.vanbuuren@tno.nl> & Karin  
Groothuis-Oudshoorn <c.g.m.oudshoorn@utwente.nl>

**Maintainer** Stef van Buuren <stef.vanbuuren@tno.nl>

**Depends** R (>= 2.10), MASS, nnet, lattice, methods

**Suggests** VIM, mitools, nlme, Zelig, lme4, survival, gamlss

**Description** Multiple Imputation using Fully Conditional Specification

**License** GPL-2 | GPL-3

**LazyLoad** yes

**LazyData** yes

**URL** <http://www.stefvanbuuren.nl>; <http://www.multiple-imputation.com>

**Repository** CRAN

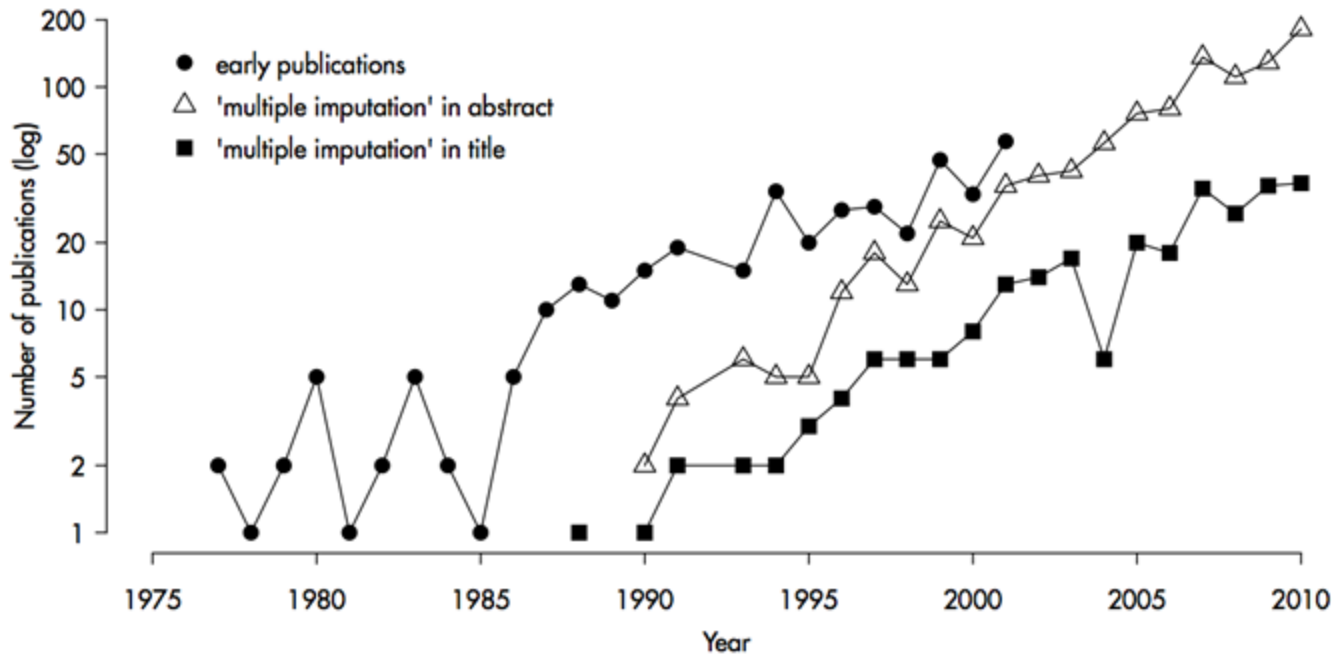
**Date/Publication** 2012-03-25 20:10:06

## R topics documented:

boys . . . . .	3
cbind.mids . . . . .	4
cc . . . . .	7
cci . . . . .	8
ccn . . . . .	9
complete . . . . .	10
fdd . . . . .	11

# www.multiple-imputation.com

[Home](#)  
[MI](#)  
[MICE](#)  
[FIMD](#)  
[Software](#)  
[Contact](#)



Number of publications (log) on multiple imputation during the period 1977–2010 according to three counting methods.

The figure contains three time series with counts of the number of publications on multiple imputation during the period 1977–2010. The search was done in Scopus on the July 11, 2011. Counts were made in three ways.

1. The right most series corresponds to the number of publications per year that featured the search term 'multiple imputation' in the title. These are often methodological articles in which new adaptations are being developed.
2. The series in the middle is the number of publication that featured 'multiple imputation' in the title, abstract or key words on the same search data. This set includes a growing group of papers that contain applications.
3. The left most series is the number of publications in a collection of early publications. This collection covers essentially everything related to multiple imputation since its inception in 1977 up to the year 2001. This group also includes chapters in books, dissertations, conference proceedings, technical reports, and so on.

Note that the vertical axis is set in the logarithm. Perhaps the most interesting series is the middle series

number of applications is growing at an exponential rate.

## CLASSIFICATION OF DOM'S

Based on Rubin (1976), Little and Rubin (1987), and Little (1995)

- **Missing completely at random (MCAR):** DOM does not depend on covariates or outcomes

$$P(r_i|x_i, y_i, \phi) = P(r_i|\phi)$$

- **Covariate-dependent (CD) missingness:** DOM may possibly depend on covariates but not outcomes

$$P(r_i|x_i, y_i, \phi) = P(r_i|x_i, \phi)$$

- **Missing at random (MAR):** DOM may depend on covariates and observed outcomes

$$P(r_i|x_i, y_i, \phi) = P(r_i|x_i, y_{i(obs)}, \phi)$$

Note that  $MCAR \subset CD \subset MAR$ .

- **Missing not at random (MNAR):** Any violation of MAR; DOM still depends on  $y_{i(mis)}$  even after any dependence on  $x_i$  and  $y_{i(obs)}$  has been accounted for

## EXPLANATION

In the case of dropout,

- MCAR means that the probability of dropout is unrelated to any characteristics of the subject at all
- CD means that the probability of dropout may be related to covariates but is unrelated to outcomes at any time
- MAR means that the probability of dropout may be related to covariates and to *pre-dropout responses*
- MNAR means that probability of dropout is related to responses at the time of dropout and possibly afterward (the latter is often not unreasonable; see Little, 1995)



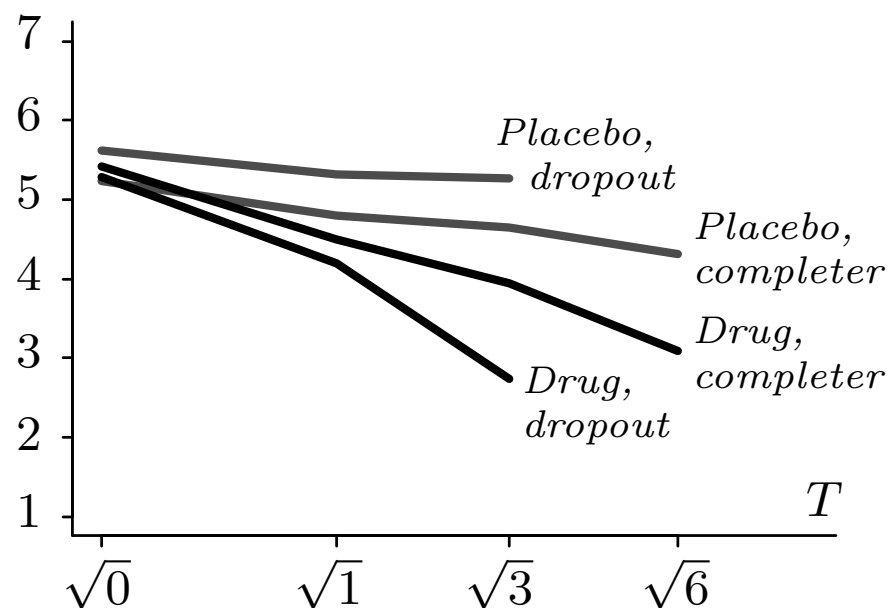
# 1. Motivation

DATA EXAMPLE FROM HEDEKER AND GIBBONS (1997)

A randomized psychiatric trial

- 312 patients received drug therapy for schizophrenia; 101 received placebo
- measurements at weeks 0, 1, 3, 6
- missing data primarily due to dropout
- outcome: severity of illness (1=normal, ..., 7=extremely ill); treat as continuous

## EXAMPLE



Based on this plot, we may conclude:

- dropout is not MCAR, because it operates differently in the treatment and control groups
- dropout is not merely CD, because completers and dropouts follow different (pre-dropout) trajectories
- dropout could be MAR or MNAR; it's impossible to tell

## 2. Basic theory

### BASIC NOTATION

$x_i$  = covariates for subject  $i$   
(assume completely observed)

$y_i$  = outcomes for subject  $i$  at all occasions  
(could be a vector or a matrix)

### THE DATA MODEL

$P(y_i|x_i, \theta)$  = some distribution

$\theta$  = population parameters of interest

For example,  $\theta$  could be

- effects of covariates on response
- difference in mean response at final occasion

Notice that  $\theta$  applies to the *entire population* of subjects

## THE MISSINGNESS

$r_i$  = binary variables indicating whether  
each element of  $y_i$  is observed or missing

- In general,  $r_i$  is a matrix of 0's and 1's of the same size as  $y_i$
- In special cases it can be reduced to a smaller set of variables
- If the only kind of missing data is dropout, then it can be reduced to a single number (time of last measurement)

## THE DISTRIBUTION OF MISSINGNESS (DOM)

$$P(r_i|x_i, y_i, \phi) = \text{some distribution}$$

First introduced by Rubin (1976, *Biometrika*); sometimes called the  
“missingness mechanism”

what we have been doing all along...

(i.e., lmer)

## Assumptions underlying this multivariate linear model analysis

- ▶ Multivariate normality, saturated means model, unstructured covariance matrix: robust for the *observed* data.
- ▶ Behaviour of the missing data (MAR):

The joint statistical behaviour of the unobserved measurements from an individual who drops out is assumed to be the same as an individual who does not dropout who shares

- ▶ the same history (i.e. previous measurements, including baseline);
- ▶ the same covariates (including treatment group).

### 3. Efficient procedures

#### A. LINEAR MIXED MODELS

- Also known as multilevel models, linear mixed-effects models, random-effects models, random-coefficient models, hierarchical linear models
- Implemented in HLM, PROC MIXED, S-PLUS, R, Stata, ...

Adopting the notation of Laird and Ware (1982), the model is

$$y_i = X_i\beta + Z_ib_i + \epsilon_i, \quad i = 1, \dots, m$$

where

$$\begin{aligned} y_i &= (y_{i1}, y_{i2}, \dots, y_{i,n_i})^T \\ b_i &\sim N_q(0, \psi) \\ \epsilon_i &\sim N_{n_i}(0, \sigma^2 V_i) \end{aligned}$$

(1)

$$\begin{aligned}\beta &= \text{fixed effects} \\ b_i &= \text{random effects for unit } i \\ \psi &= \text{between-unit covariance matrix} \\ \sigma^2 V_i &= \text{within-unit covariance matrix}\end{aligned}$$

- Handles unequal  $n_i$ 's, time-varying covariates, unequally spaced responses
- Often we use  $V_i = I$ , but other structures—e.g., autoregressive—are useful, especially when  $n_i$ 's are large
- measurement times are often incorporated into  $X_i$ ,  $Z_i$  as polynomials
- $Z_i$  contains a subset of the columns of  $X_i$

An excellent treatment these models is the new book by Fitzmaurice, Laird and Ware (2004)

## WHAT ABOUT MISSING DATA?

1. When data are unbalanced by design, then ML or REML estimation is the right thing to do
2. If some responses for some subjects are missing, we may omit the missed occasions and apply ML or REML to the reduced data; this is appropriate if the missing responses are MAR
3. Note that important correlates of missingness need to be included in the model for MAR to be plausible

## NOTES ABOUT PROC MIXED

(This is not necessarily limited to PROC MIXED; other programs may behave in a similar fashion)

- PROC MIXED will automatically omit occasions with missing responses (which is good under MAR)
- PROC MIXED will also omit subjects or occasions with missing **covariates**, which implicitly assumes that these are MCAR (not so good)



TABLE 3.

### Summary of Available Options for Handling Missing Data

Action	Pros	Cons	Recommendation
Careful deliberation about why data are missing	Helps determine correct model	Not always clear	Always do this
Last observation carried forward	Easy to do	Unrealistic assumption, overestimates precision	Never do this
Mean imputation	Easy to do, preserves mean	Does not preserve relationships in data. Overestimates precision.	Never do this
Complete-case analysis	Easy to do	Biased estimates unless data are MCAR. Loss of information.	Never do this.
End-point analysis	Missing values no longer an issue	Ignores information, ignores time	Never do this.
Single imputation	Reduces bias	Overestimates precision.	Use multiple imputation.
Mixed-effects regression model	Makes use of all available information	Can be complicated to fit. Assumes MAR.	Often a good choice
Multiple imputation	Allows one to incorporate auxiliary variables into imputation model	Requires expertise. Additional steps for analyzing data.	Do it if MAR assumption is likely to be satisfied
Nonignorable models	Explore the effect of different missing data assumptions	Not clear what is correct model. Can be complicated.	Worth doing, especially as a sensitivity analysis
Fit several different statistical models that make different assumptions regarding why data are missing	Sensitivity of inferences to different assumptions regarding the missing data mechanism.	Additional work. May complicate the overall picture.	Worth doing.

### Lavori, part B

the study should be included in the imputation- ML, irrespective of the inclusion of dosage to adequately handle missing data may

## AD-HOC (AND GENERALLY FLAWED) APPROACHES FOR HANDLING MISSING DATA

We describe here the common ad-hoc approaches for handling missing values, which are often used when analyzing longitudinal data, because they are easy to implement and do not require special software. Despite their common use, they rely on implicit assumptions that are usually unreasonable and often lead to invalid inference.

### Last observation carried forward

With last observation carried forward (LOCF), missing values are replaced with the most recent previously observed value in the same patient. The filled in dataset is then analyzed as if there had been no missing data. This substitution of previously observed values for missing data can be performed for both intermittent missing values and measurement dropouts in repeated measures designs. Very strong and often unrealistic assumptions have to be made to ensure the validity of this method. First, LOCF assumes that a subject's true but unmeasured status stays at the same level from the moment of truncation onward (or during the period they are unobserved in the case of intermittent missingness).<sup>7</sup> In other words, there is a perfect relationship between the last observation and those following it. The prior trajectory of the subject is not taken into account, and any change is assumed to level off immediately. For intermittent missing data, the subsequent trajectory of the subject after the "gap" is not taken into account either. Further, as will be discussed later in this article, LOCF (like all substitution and single imputation procedures) overestimates precision by treating imputed and actually observed values on equal footing. It is often believed, erroneously, that LOCF is conservative, thus does not lead to an inflated type I error rate. For point estimates, LOCF might underestimate the

improvement in the experimental arm, if there is a systematic improvement in the outcome over time. However, the same underestimation might also happen in the control/placebo arm. Therefore, it is not clear whether the treatment effect based on contrasting the trajectories in the two arms is under-estimated or not. Furthermore, the overestimation of the precision might lead to underestimation of the standard error and inflation of the type I error. There are several published examples where LOCF does poorly.<sup>8-10</sup>

### Mean substitution

In the context of longitudinal studies, mean substitution is typically implemented by replacing a missing value with the average (over other patients') observed value for the same variable and then analyzing the dataset as if it were complete. Although this method does preserve the overall mean for the time period, it has two serious disadvantages. Mean substitution does not preserve relationships among other variables in the data. For example, if a subject's month 2 depression score is missing, substitution of the mean at month 2 ignores that person's depression score for months 1 and 3. Mean substitution, therefore, always attenuates correlations between the measures. Finally, as with all substitution and single imputation procedures, mean substitution does not take into account uncertainty in the true but unknown value.

### Regression substitution

Regression substitution extends the mean substitution method by using a regression substitution estimate to replace a missing data point. For each subject's missing data, the predictor variables consist of all those that are non-missing, with regression substitution coefficients computed from the remaining data. Although this procedure is a substantial improvement over LOCF and mean substitution, it is still unsatisfactory because

missing data are replaced with values having too little variability, resulting in bias in correlations and over-estimation of the precision.

### Complete-case analysis

Complete-case analysis involves discarding all observed data elements for subjects who have any missing values and restricting the data analysis to the remaining complete cases. This is the simplest procedure for handling missing data. It is usually done automatically by most software packages when missing data are encountered so that the dataset can be analyzed using standard complete-data methods. Unless the observations with missing values are only randomly different from those without missing values (ie, unless the data are MCAR), complete-case analysis will produce biased estimates. Complete-case analysis can also result in substantial information loss, by discarding an entire subject's data because of a few missing items. Rather than discarding an entire observation because of a single missing value, methods that make better use of all available information will provide estimates that are more precise and less biased.

### End-point analysis

End-point analysis, a form of LOCF (see Gibbons and colleagues<sup>11</sup> for a review of limitations) is a procedure that concentrates on baseline and the last observed measurement for each individual, ignoring all observations between these times. Although the baseline period is usually the same for each individual, the end point will be different for each individual depending on if and when they drop out of the study. Typically, some form of difference or adjusted score is calculated from the baseline and end-point scores, and these difference or adjusted scores are compared across treatment groups.

By using only the last observed measurement for each individual, missing

values are no longer an issue (except for those who have no follow-up data). However, there are many drawbacks to this approach. First, data between the first and last time points are ignored. This is problematic because a large amount of information is being discarded leading to reduced efficiency of parameter estimates. In addition, the researcher is no longer able to study individual trends over time, one of the original goals of longitudinal research.

A further drawback to end-point analysis is that since the time of the last measurement can vary for each individual, time is effectively ignored in the analysis. As a result, between-group comparisons can be confounded with time, since subjects in one group may have been assessed under a different period than subjects from another group. Within each group the length of the period itself may be influenced by the treatment. For example, if placebo-treated participants are more likely to drop out earlier than participants receiving the active drug, estimates of the treatment effect will favor the active drug even if the improvement rate is identical.<sup>12</sup>

### Single imputation

Single imputation is a general method of replacing missing values with plausible values. It differs from the previous methods in that the imputed value has the same distribution as the non-missing data. One way to do this is to correct the regression substitution method, which uses a prediction equation to adjust for a person's own non-missing variables by adding in a random component to mimic the additional variability that real data would be expected to have around this predicted value. For each variable that has any missing data, a regression substitution model for imputation is developed, which uses a person's non-missing data to form a best predictor of that person's missing data. To this predictor, a random component is added based on

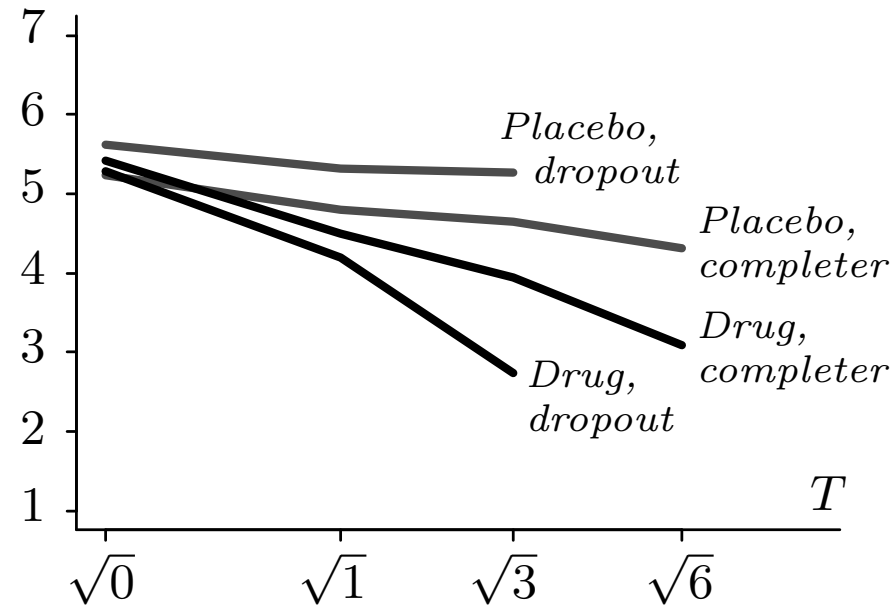
# 1. Motivation

DATA EXAMPLE FROM HEDEKER AND GIBBONS (1997)

A randomized psychiatric trial

- 312 patients received drug therapy for schizophrenia; 101 received placebo
- measurements at weeks 0, 1, 3, 6
- missing data primarily due to dropout
- outcome: severity of illness (1=normal, ..., 7=extremely ill); treat as continuous

Plot of average response versus square root of week



A completers-only analysis would severely understate the treatment effect. We want a sensible procedure to analyze the incomplete data

- low bias
- high efficiency
- robust to assumptions about from population distribution and missing-data mechanisms

## CASE-DELETION METHODS

Often used in the past to produce balanced datasets for repeated-measures ANOVA

- Delete any subject with a missing value at any occasion
- Perhaps delete some complete subjects as well to balance the  $n$ 's across treatment groups

Modern methods for analyzing longitudinal data (e.g. PROC MIXED) do not require balance, so case-deletion procedures have become less popular

## A few comments on case deletion

- Not so bad for laboratory experiments, for which data are often nearly balanced
- In studies with human subjects (especially over longer periods of time), missed measurements and dropout are a more serious issue
- When completers and dropouts seem to follow different trajectories, analyzing only the completers may be very misleading
- For population inferences, it's nearly always better to analyze the data from all subjects whether they completed the study or not
  - less biased
  - more efficient

## MEAN IMPUTATION

$y_{ij}$  = response for subject  $i$  at occasion  $j$

$r_{ij}$  = 1 if  $y_{ij}$  observed, 0 if missing

If  $y_{ij}$  is missing, we can replace it by

- the mean response for subject  $i$

$$y_{i\cdot} = \frac{\sum_j r_{ij} y_{ij}}{\sum_j r_{ij}}$$

- the mean response for occasion  $j$

$$y_{\cdot j} = \frac{\sum_i r_{ij} y_{ij}}{\sum_i r_{ij}}$$

Both of these methods may seriously distort estimates and measures of uncertainty



## LAST OBSERVATION CARRIED FORWARD

For attrition (dropout): If a subject drops out after occasion  $j$ ,

replace  $y_{i,j+1}, y_{i,j+2}, \dots$  by  $y_{i,j}$

- Equivalent to subject-mean imputation for dropout after first occasion
- Tends to understate differences in estimated time-trends between treatment and control groups (thought to be “conservative”)
- Not necessarily “conservative,” because standard errors are biased downward as well
- Especially bad for outcomes that have high variation within a subject

## Simple ad hoc methods: LOCF

### Last Observation Carried Forward (LOCF):

- ▶ Following dropout, an individual's last measurement is imputed as a replacement for all the subsequent missing observations.
- ▶ Sometimes even a baseline measurement is carried forward in the same way.

Analysis of change from baseline then implies a zero is always imputed.

- ▶ LOCF is often proposed as the primary analysis in longitudinal clinical trials.
- ▶ It is often combined with a completers analysis as a form of “sensitivity analysis”
- ▶ **FDA** Guidance for Industry (from the website):

*The problem of dropouts is not resolved by an intention-to-treat (...) analysis with an imputation by last observation carried forward.*

- ▶ LOCF and other “simple” imputation methods are not *principled*.
- ▶ LOCF can create treatment effects when none exist, and mask real effects.
- ▶ The assumptions under which LOCF is valid are contrived and unrealistic.
- ▶ If such assumptions are to be made they can be incorporated into principled analyses.
- ▶ In missing data settings simple analyses rarely imply simple assumptions.

## COMMENTS ON IMPUTATION IN GENERAL

- Single-imputation strategies designed to precisely predict the missing values tend to distort estimates of population quantities
- The goal of the missing-data procedure is to draw accurate inferences about population quantities (e.g. mean change over time), not to accurately predict the missing values
- With imputation, the best way to achieve that goal is to preserve all aspects of the data distribution (means, trends, within- and between-subject variation, etc.)
- Ad hoc imputation methods inevitably preserve some aspects but distort others

# Chapter 9

---

## *Longitudinal data*

---

### 9.1 Long and wide format

Longitudinal data can be coded into “long” and “wide” formats. A wide dataset will have one record for each individual. The observations made at different time points are coded as different columns. In the wide format every measure that varies in time occupies a set of columns. In the long format there will be multiple records for each individual. Some variables that do not vary in time are identical in each record, whereas other variables vary across the records. The long format also needs a “time” variable that records the time in each record, and an “id” variable that groups the records from the same person.

A simple example of the wide format is

```
id age Y1 Y2
1  14 28 22
2  12 34 16
3  ...
```

In the long format, this dataset looks like

```
id age Y
1  14 28
1  14 22
2  12 34
2  12 16
3  ...
```

Note that the concepts of long and wide are general, and also apply to cross-sectional data. For example, we have seen the long format before in Section 6.1.1, where it referred to stacked imputed data that was produced by the `complete()` function. The basic idea is the same.

Both formats have their advantages. If the data are collected on the same time points, the wide format has no redundancy or repetition. Elementary statistical computations like calculating means, change scores, age-to-age correlations between time points, or the  $t$ -test are easy to do in this format. The

long format is better at handling irregular and missed visits. Also, the long format has an explicit time variable available that can be used for analysis. Graphs and statistical analyses are easier in the long format.

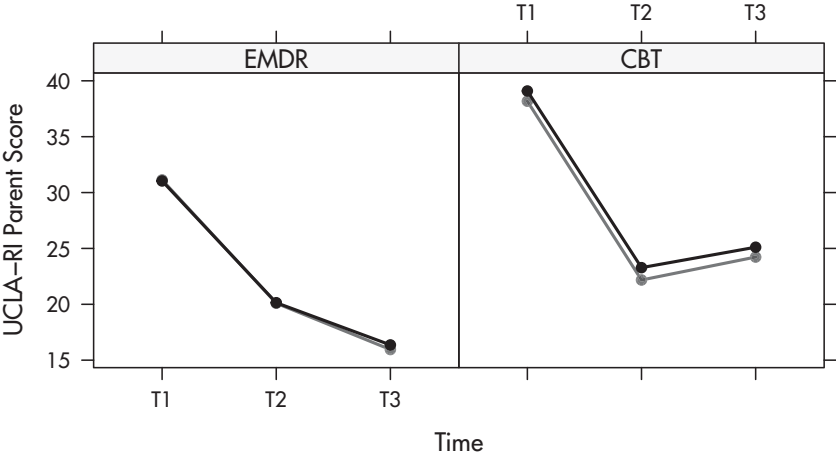
Applied researchers often collect, store and analyze their data in the wide format. Classic ANOVA and MANOVA techniques for repeated measures and structural equation models for longitudinal data assume the wide format. Modern multilevel techniques and statistical graphs, however, work only from the long format. The distinction between the two formats is a first stumbling block for those new to longitudinal analysis.

Singer and Willett (2003) advise the data storing in both formats. The wide and the long formats can be converted into each other by a database operation. R and Stata have `reshape()` functions. In SPSS the wide-to-long conversion is done by the `VARSTOCASES` commands, and the long-to-wide conversion by `CASESTOVAR`s. Both are available from the `Data Restructure...` menu. SAS uses `PROC TRANSPOSE` for this purpose.

Multiple imputation of longitudinal data is conveniently done when data are in the wide format. Apart from the fact that the columns are ordered in time, there is nothing special about the imputation problem. We may thus apply the techniques from the earlier chapters to longitudinal data. Section 9.2 discusses an imputation technique in the wide format in a clinical trial application with the goal of performing a statistical analysis according to the intention to treat (ITT) principle. The longitudinal character of the data helped specify the imputation model.

The wide-to-long conversion can usually be done without a problem. The long-to-wide conversion can be difficult. If individuals are seen at different times, direct conversion is impractical. The number of columns in the wide format becomes overly large, and each column contains many missing values. An ad hoc solution is to create homogeneous time groups, which then become the new columns in the wide format. Such regrouping will lead to loss of precision of the time variable. For some studies this need not be a problem, but for others it will.

A more general approach is to impute data in the long format, which requires some form of multilevel imputation. Section 9.3 discusses multiple imputation in the long format. The application defines a common time raster for all persons. Multiple imputations are drawn for each raster point. The resulting imputed datasets can be converted to, and analyzed in, the wide format if desired. This approach is a more principled way to deal with the information loss problem discussed previously. The procedure aligns times to a common raster, hence the name *time raster imputation* (cf. Section 9.3).



**Figure 9.2:** Mean levels of PTSD-RI Parent Form for the completely observed profiles (gray) and all profiles (black) in the EMDR and CBT groups.

presumably caused by the EMDR and CBT therapies. The shape between end of treatment (T2) and follow-up (T3) differs somewhat for the group, suggesting that EMDR has better long-term effects, but this difference was not statistically significant. Also note that the complete case analysis and the analysis based on ITT are in close agreement with each other here.

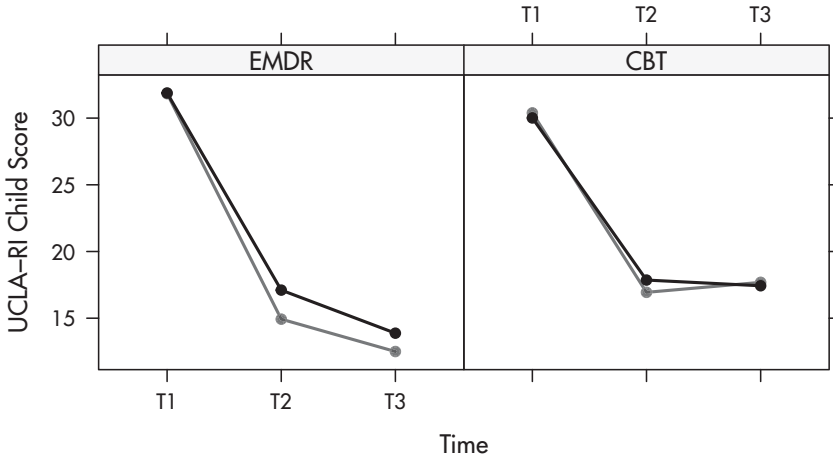
We will not go into details here to answer the second research question as stated on p. 223. It is of interest to note that EMDR needed fewer sessions to achieve its effect. The original publication (De Roos et al., 2011) contains the details.

### 9.3 Time raster imputation

Longitudinal analysis has become virtually synonymous with mixed effects modeling. Following the influential work of Laird and Ware (1982) and Jennrich and Schluchter (1986), this approach characterizes individual growth trajectories by a small number of random parameters. The differences between individuals are expressed in terms of these parameters.

In some applications, it is natural to consider *change scores*. Change scores are however rather awkward within the context of mixed effects models. This section introduces *time raster imputation*, a new method to generate imputations on a regular time raster from irregularly spaced longitudinal data.





**Figure 9.3:** Mean levels of PTSD-RI Child Form for the completely observed profiles (gray) and all profiles (black) in the EMDR and CBT groups.

The imputed data can then be used to calculate change scores or age-to-age correlations, or apply quantitative techniques designed for repeated measures.

### 9.3.1 Change score

Let  $Y_1$  and  $Y_2$  represent repeated measurements of the same object at times  $T_1$  and  $T_2$  where  $T_1 < T_2$ . The difference  $\Delta = Y_2 - Y_1$  is the most direct measure of change over time. Willett (1989, p. 588) characterized the change score as an “intuitive, unbiased, and computationally-simple measure of individual growth.”

One would expect that modern books on longitudinal data would take the change score as their starting point. That is not the case. The change score is fully absent from most current books on longitudinal analysis. For example, there is no entry “change score” in the index of Verbeke and Molenberghs (2000), Diggle et al. (2002), Walls and Schafer (2006) or Fitzmaurice et al. (2009). Singer and Willett (2003, p. 10) do discuss the change score, but they quickly dismiss it on the basis that a study with only two time points cannot reveal the shape of a person’s growth trajectory.

The change score, once the centerpiece of longitudinal analysis, has disappeared from the methodological literature. I find this is somewhat unfortunate as the parameters in the mixed effects model are more difficult to interpret than the change score. Moreover, classic statistical techniques, like the paired  $t$ -test or split-plot ANOVA, are built on the change score. There is a gap be-

tween modern mixed effects models and classical linear techniques for change scores and repeated measures data.

Calculating a mean change score is only sensible if different persons are measured at the same time points. When the data are observed at irregular times, there is no simple way to calculate change scores. Calculating change scores from the person parameters of the mixed effects model is technically trivial, but such scores are difficult to interpret. The person parameters are fitted values that have been smoothed. Deriving a change score as the difference between the fitted curve of the person at  $T_1$  and  $T_2$  results in values that are closer to zero than those derived from data that have been observed.

This section describes a technique that inserts pseudo time points to the observed data of each person. The outcome data at these supplementary time points are multiply imputed. The idea is that the imputed data can be analyzed subsequently by techniques for change scores and repeated measures.

The imputation procedure is akin to the process needed to print a photo in a newspaper. The photo is coded as points on a predefined raster. At the microlevel there could be information loss, but the scenery is essentially unaffected. Hence the name *time raster imputation*. My hope is that this method will help bridge the gap between modern and classic approaches to longitudinal data.

### 9.3.2 Scientific question: Critical periods

The research was motivated by the question: *At what ages do children become overweight?* Knowing the answer to this question may provide handles for preventive interventions to counter obesity.

Dietz (1994) suggested the existence of three *critical periods* for obesity at adult age: the prenatal period, the period of adiposity rebound (roughly around the age of 5–6 years), and adolescence. Obesity that begins at these periods is expected to increase the risk of persistent obesity and its complications. Overviews of studies on critical periods are given by Cameron and Demerath (2002) and Lloyd et al. (2010).

In the sequel, we use the body mass index (BMI) as a measure of overweight. BMI will be analyzed in standard deviation scores (SDS) using the relevant Dutch references (Fredriks et al., 2000a,b). Our criterion for being overweight in adulthood is defined as BMI SDS  $\geq 1.3$ .

As an example, imagine an 18-year old person with a BMI SDS equal to +1.5 SD. How did this person end up at 1.5 SD? If we have the data, we can plot the measurements against age, and study the individual track. The BMI SDS trajectory may provide key insights into development of overweight and obesity.

Figure 9.4 provides an overview of five theoretical BMI SDS trajectories that the person might have followed. These are:

1. *Long critical period.* A small but persistent centile crossing across the entire age range. In this case, everything (or nothing) is a critical period.



---

# *Journal of Statistical Software*

December 2011, Volume 45, Issue 3.

<http://www.jstatsoft.org/>

---

## **mice: Multivariate Imputation by Chained Equations in R**

Stef van Buuren  
TNO

Karin Groothuis-Oudshoorn  
University of Twente

---

### Abstract

The R package **mice** imputes incomplete multivariate data by chained equations. The software **mice** 1.0 appeared in the year 2000 as an S-PLUS library, and in 2001 as an R package. **mice** 1.0 introduced predictor selection, passive imputation and automatic pooling. This article documents **mice** 2.9, which extends the functionality of **mice** 1.0 in several ways. In **mice** 2.9, the analysis of imputed data is made completely general, whereas the range of models under which pooling works is substantially extended. **mice** 2.9 adds new functionality for imputing multilevel data, automatic predictor selection, data handling, post-processing imputed values, specialized pooling routines, model selection tools, and diagnostic graphs. Imputation of categorical data is improved in order to bypass problems caused by perfect prediction. Special attention is paid to transformations, sum scores, indices and interactions using passive imputation, and to the proper setup of the predictor matrix. **mice** 2.9 can be downloaded from the Comprehensive R Archive Network. This article provides a hands-on, stepwise approach to solve applied incomplete data problems.

*Keywords:* MICE, multiple imputation, chained equations, fully conditional specification, Gibbs sampler, predictor selection, passive imputation, R.

---

## 1. Introduction

Multiple imputation (Rubin 1987, 1996) is the method of choice for complex incomplete data problems. Missing data that occur in more than one variable presents a special challenge. Two general approaches for imputing multivariate data have emerged: joint modeling (JM) and fully conditional specification (FCS), also known as multivariate imputation by chained equations (MICE). Schafer (1997) developed various JM techniques for imputation under the multivariate normal, the log-linear, and the general location model. JM involves specifying a multivariate distribution for the missing data, and drawing imputation from their conditional

	1	2	3	4	5
1	20.4	27.2	22.0	25.5	27.4
3	27.4	22.5	24.9	22.7	33.2
4	20.4	20.4	24.9	27.2	27.5
6	22.5	27.5	26.3	20.4	24.9
10	27.2	20.4	27.2	26.3	22.7
11	22.7	22.5	22.7	29.6	25.5
12	29.6	28.7	22.5	33.2	27.4
16	27.4	22.5	35.3	22.7	20.4
21	30.1	27.4	24.9	20.4	27.2

Imputations for `bmi` are now sampled (by `mice.impute.pmm()`) under the intercept-only model. Note that these imputations are appropriate only under the MCAR assumption.

### *Multilevel imputation*

Imputation of multilevel data poses special problems. Most techniques have been developed under the joint modeling perspective (Schafer and Yucel 2002; Yucel 2008; Goldstein *et al.* 2009). Some work within the context of FCS has been done (Jacobusse 2005), but this is still an open research area. The `mice` 2.9 package include the `mice.impute.2L.norm()` function, which can be used to impute missing data under a linear multilevel model. The function was contributed by Roel de Jong, and implements the Gibbs sampler for the linear multilevel model where the within-class error variance is allowed to vary (Kasim and Raudenbush 1998). Heterogeneity in the variances is essential for getting good imputations in multilevel data. The method is an improvement over simpler methods like flat-file imputation or per-group imputation (van Buuren 2010).

Using `mice.impute.2L.norm()` (or equivalently `mice.impute.2l.norm()`) deviates from other univariate imputation functions in `mice` 2.9 in two respects. It requires the specification of the fixed effects, the random effects and the class variable. Furthermore, it assumes that the predictors contain a column of ones representing the intercept. Random effects are coded in the predictor matrix as a '2'. The class variable (only one is allowed) is coded by a '-2'. The example below uses the popularity data of (Hox 2002). The dependent variable is `pupil popularity`, which contains 848 missing values. There are two random effects: `const` (intercept) and `sex` (slope), one fixed effect, teacher experience (`texp`), and one class variable (`school`). Imputations can be generated as

```
R> popmis[1:3, ]
```

	pupil	school	popular	sex	texp	const	teachpop
1	1	1	NA	1	24	1	7
2	2	1	NA	0	24	1	7
3	3	1	7	1	24	1	6

```
R> ini <- mice(popmis, maxit = 0)
R> pred <- ini$pred
R> pred["popular", ] <- c(0, -2, 0, 2, 1, 2, 0)
R> imp <- mice(popmis, meth = c("", "", "2l.norm", "", ""),
+           "", ""), pred = pred, maxit = 1, seed = 71152)
```

## References

White, I. R., Royston, P. (2009). Imputing missing covariate values for the Cox model. *Statistics in Medicine*, 28(15), 1982-1998.

van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Boca Raton, FL: Chapman & Hall/CRC Press.

## Examples

```
leuk$status <- 1 ## no censoring occurs in leuk data (MASS)
ch <- nelsonaalen(leuk, time, status)
plot(x = leuk$time, y = ch, ylab="Cumulative hazard", xlab="Time")

### See example on http://www.engineeredsoftware.com/lmar/pe_cum_hazard_function.htm
time <- c(43, 67, 92, 94, 149, rep(149,7))
status <- c(rep(1,5),rep(0,7))
eng <- data.frame(time, status)
ch <- nelsonaalen(eng, time, status)
plot(x = time, y = ch, ylab="Cumulative hazard", xlab="Time")
```

---

nhanes

NHANES example - all variables numerical

---

## Description

A small data set with non-monotone missing values.

## Usage

```
data(nhanes)
```

## Format

A data frame with 25 observations on the following 4 variables.

**age** Age group (1=20-39, 2=40-59, 3=60+)

**bmi** Body mass index (kg/m\*\*2)

**hyp** Hypertensive (1=no,2=yes)

**chl** Total serum cholesterol (mg/dL)

## Details

A small data set with all numerical variables. The data set nhanes2 is the same data set, but with age and hyp treated as factors.



---

# *Journal of Statistical Software*

December 2011, Volume 45, Issue 3.

<http://www.jstatsoft.org/>

---

## **mice: Multivariate Imputation by Chained Equations in R**

Stef van Buuren  
TNO

Karin Groothuis-Oudshoorn  
University of Twente

---

### Abstract

The R package **mice** imputes incomplete multivariate data by chained equations. The software **mice** 1.0 appeared in the year 2000 as an S-PLUS library, and in 2001 as an R package. **mice** 1.0 introduced predictor selection, passive imputation and automatic pooling. This article documents **mice** 2.9, which extends the functionality of **mice** 1.0 in several ways. In **mice** 2.9, the analysis of imputed data is made completely general, whereas the range of models under which pooling works is substantially extended. **mice** 2.9 adds new functionality for imputing multilevel data, automatic predictor selection, data handling, post-processing imputed values, specialized pooling routines, model selection tools, and diagnostic graphs. Imputation of categorical data is improved in order to bypass problems caused by perfect prediction. Special attention is paid to transformations, sum scores, indices and interactions using passive imputation, and to the proper setup of the predictor matrix. **mice** 2.9 can be downloaded from the Comprehensive R Archive Network. This article provides a hands-on, stepwise approach to solve applied incomplete data problems.

*Keywords:* MICE, multiple imputation, chained equations, fully conditional specification, Gibbs sampler, predictor selection, passive imputation, R.

---

## 1. Introduction

Multiple imputation (Rubin 1987, 1996) is the method of choice for complex incomplete data problems. Missing data that occur in more than one variable presents a special challenge. Two general approaches for imputing multivariate data have emerged: joint modeling (JM) and fully conditional specification (FCS), also known as multivariate imputation by chained equations (MICE). Schafer (1997) developed various JM techniques for imputation under the multivariate normal, the log-linear, and the general location model. JM involves specifying a multivariate distribution for the missing data, and drawing imputation from their conditional

(1997, p. 237). The data contains four variables: `age` (age group), `bmi` (body mass index), `hyp` (hypertension status) and `chl` (cholesterol level). The data are stored as a `data frame`. Missing values are represented as `NA`.

```
R> nhanes
```

```
   age  bmi hyp chl
1    1   NA NA  NA
2    2 22.7  1 187
3    1   NA  1 187
4    3   NA NA  NA
5    1 20.4  1 113
... 
```

### *Inspecting the missing data*

The number of the missing values can be counted and visualized as follows:

```
R> md.pattern(nhanes)
```

```
   age hyp bmi chl
13   1   1   1   1  0
  1   1   1   0   1  1
  3   1   1   1   0  1
  1   1   0   0   1  2
  7   1   0   0   0  3
    0   8   9  10 27
```

There are 13 (out of 25) rows that are complete. There is one row for which only `bmi` is missing, and there are seven rows for which only `age` is known. The total number of missing values is equal to  $(7 \times 3) + (1 \times 2) + (3 \times 1) + (1 \times 1) = 27$ . Most missing values (10) occur in `chl`.

Another way to study the pattern involves calculating the number of observations per patterns for all pairs of variables. A pair of variables can have exactly four missingness patterns: both variables are observed (pattern `rr`), the first variable is observed and the second variable is missing (pattern `rm`), the first variable is missing and the second variable is observed (pattern `mr`), and both are missing (pattern `mm`). We can use the `md.pairs()` function to calculate the frequency in each pattern for all variable pairs as

```
R> p <- md.pairs(nhanes)
```

```
R> p
```

```
$rr
```

```
   age bmi hyp chl
age  25  16  17  15
bmi  16  16  16  13
```

```
# refer to van Buuren S and Groothuis-Oudshoorn K (2011). mice: Multivariate Imputation by Chained
Equations in R. Journal of Statistical Software, 45(3), 1-67. http://www.jstatsoft.org/v45/i03/
R version 2.14.1 (2011-12-22)
> install.packages("mice") # imputation package
> install.packages("VIM") # additional graphics
> library(mice)
> library(VIM)
```

```
#number of observations per patterns for all pairs of variables
> data(nhanes)
> head(nhanes)
  age  bmi  hyp chl
1    1   NA   NA  NA
2    2 22.7    1 187
3    1   NA    1 187
4    3   NA   NA  NA
5    1 20.4    1 113
6    3   NA   NA 184

> md.pattern(nhanes) #age complete
  age hyp bmi chl
13    1    1    1    0
   1    1    0    1    1
   3    1    1    0    1
   1    1    0    0    1    2
   7    1    0    0    0    3
   0    8    9   10   27
```

```
#multiply imputed data set is stored in the object impdrr of class mids
```

```
> impdrr = mice(nhanes, seed = 23109) # m=5 imps is default
iter imp variable
  1    1  bmi  hyp  chl
  1    2  bmi  hyp  chl
  1    3  bmi  hyp  chl
  .....
  5    3  bmi  hyp  chl
  5    4  bmi  hyp  chl
  5    5  bmi  hyp  chl #default method, numerical data, predictive mean matching (pmm)
```

```
#The complete() function extracts the five imputed data sets from the imp object as a
long (row-stacked) matrix with 125 records
```

```
> stripplot(impdrr, pch = 20, cex = 1.2)
```

```
#The fit object has class mira and contains the results of 5 complete-data analyses
```

```
> fitdrr = with(impdrr, lm(chl ~ age + bmi))
```

```
# pool separate results
```

```
> round(summary(pool(fitdrr)), 2) # match Stef JSS results
```

	est	se	t	df	Pr(> t )	lo 95	hi 95	nmis	fmi	lambda
(Intercept)	-34.16	76.07	-0.45	6.81	0.67	-215.05	146.73	NA	0.57	0.47
age	34.33	14.86	2.31	4.04	0.08	-6.76	75.42	0	0.75	0.65
bmi	6.21	2.21	2.81	8.80	0.02	1.20	11.23	9	0.48	0.37

```
> impdrr2 = mice(nhanes, seed = 23009) #try again or set m=50, set.seed
```

```
> fitdrr2 = with(impdrr2, lm(chl ~ age + bmi))
```

```
> round(summary(pool(fitdrr2)), 2) #this is not closer to JSS, diff seed
```

	est	se	t	df	Pr(> t )	lo 95	hi 95	nmis	fmi	lambda
(Intercept)	-32.13	59.67	-0.54	14.04	0.60	-160.08	95.82	NA	0.29	0.20
age	35.76	10.50	3.41	8.52	0.01	11.80	59.72	0	0.49	0.38
bmi	6.12	1.87	3.28	14.86	0.01	2.14	10.11	9	0.27	0.17

```
>
```



the B-K class example (posted). Finally use all 4 longitudinal measures (weeks 0,1,4,6) for a Active vs Placebo comparison using lmer. Compare with the results that use only 2 observations.

3. Crossover Design. The dataset consists of safety data from a crossover trial on the disease cerebrovascular deficiency. The response variable is not a trial endpoint but rather a potential side effect. In this two-period crossover trial, comparing the effects of active drug to placebo, 67 patients were randomly allocated to the two treatment sequences, with 34 patients receiving placebo followed by active treatment, and 33 patients receiving active treatment followed by placebo. The response variable is binary, indicating whether an electrocardiogram (ECG) was abnormal ( $Y=1$ ) or normal ( $Y=0$ ). Each patient has a bivariate binary response vector.

Data set is available at <http://www.hsph.harvard.edu/fitzmaur/ala/ecg.txt> (needs to be cut-and-paste into editor). Carry out the basic analysis of variance for this crossover design following week 5 Lecture topic 2. You may want to use glm to take into account the binary outcome. Does the treatment increase the probability of abnormal ECG? Give a point estimate and significance test for the treatment effect.

4. Data on Amenorrhea from Clinical Trial of Contracepting Women. Source: Table 1 (page 168) of Machin et al. (1988). Reference: Machin D, Farley T, Busca B, Campbell M and d'Arcangues C. (1988). Assessing changes in vaginal bleeding patterns in contracepting women. *Contraception*, 38, 165-179.

Data in long form and a wide-form version

Description: The data are from a longitudinal clinical trial of contracepting women. In this trial women received an injection of either 100 mg or 150 mg of depot-medroxyprogesterone acetate (DMPA) on the day of randomization and three additional injections at 90-day intervals. There was a final follow-up visit 90 days after the fourth injection, i.e., one year after the first injection. Throughout the study each woman completed a menstrual diary that recorded any vaginal bleeding pattern disturbances. The diary data were used to determine whether a woman experienced amenorrhea, the absence of menstrual bleeding for a specified number of days. A total of 1151 women completed the menstrual diaries and the diary data were used to generate a binary sequence for each woman according to whether or not she had experienced amenorrhea in the four successive three month intervals.

In clinical trials of modern hormonal contraceptives, pregnancy is exceedingly rare (and would be regarded as a failure of the contraceptive method), and is not the main outcome of interest in this study. Instead, the outcome of interest is a binary response indicating whether a woman experienced amenorrhea in the four successive three month intervals. A feature of this clinical trial is that there was substantial dropout. More than one third of the women dropped out before the completion of the trial. In the linked data, missing data are designated by "." [note: in the week 6 terminology consider the dropouts to be *missing at random*, not necessarily a correct assumption.]

The purpose of this analysis is to assess the influence of dosage on the risk of amenorrhea and any individual differences in the risk of amenorrhea.

Show your model for these data and the results. Provide significance tests and/or interval estimates for the odds of amenorrhea as a function of dose. Display and interpret individual differences in response by showing the random effects within each experimental group.

5. Chick Data, *finale*. One more use of the chick data (week 3, problem 2; week 1 class lecture). Use the data for all 4 Diets to construct a nlmer model that allows asymptotes to differ across the four diets. Do the diets produce significantly different results? Which diet produces the heaviest 'mature' chick weight?

6. Missing Data. Wide-form longitudinal data

Artificial data example from week 2 RQ3 and Week 4 Lecture item 4 (used in Myths examples to illustrate time-1, time-2 data analysis) [Two part artificial data example](#). The top frame (the Xi's) is 40 subjects each with three equally spaced time observations (here in wide form). For these these perfectly measured "Xi" measurements each subject's observation fall on a straight-line.

a. Use data set [W6prob1a](#), for which about 15% of the observations have been made missing. Use these data (with lm) to recreate the multiple regression demonstration in Week 4 lecture, part 4: "Correlates and predictors of change: time-1, time-2 data". Compare with the results for the full data on 40 subjects. What does lm do with missing data?

b. Repeat part a with data set [W6prob1b](#). Can you find any reason to doubt a "missing at random" assumption for this data set?

Note: in Week 10 we will demonstrate multiple imputation procedures (mice) for wide-form data, at least.

7. *Beat the Blues* from Chap 12 of HSAUR 2nd ed (resource # 4).

Data in wide form: `data("BtheB", package = "HSAUR2")`. Chap. 12 describes the cognitive behavioural program and conducts various analyses. We will use the pretest and the two-month followup (additional followups have lots of missing data).

Investigate the effectiveness of Beat the Blues from these 2-wave data.

## 11/2. Comparing Group Growth, continued. Observational Studies, Cohort Designs.

### Longitudinal in the news

Another crossover design (from Stat266). RCT (cross-over design). [Damn right! The secret of success is swearing: How shouting four letter words can help make you stronger](#) [Swearing can help you boost your physical performance](#) [The full power of swearing is starting to be discovered](#)

### Lecture Topics

#### Week 6

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](#) Publication: 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](#)

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, minimum wage ex](#)

R-package `wfe` (my failures). paper [On the Use of Linear Fixed Effects Regression Models for Causal Inference](#) (sec 3.2)

D. Interrupted time-series.

Intros: [Interrupted Time Series Quasi-Experiments](#) Gene V Glass Arizona State University.

[Time Series Analysis with R](#) section 4.6 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. 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. [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

## lmer, missing data

```
> ncM = read.table("D:\\drr17\\stat222\\week10\\ncLong_dataM", header = T)
> head(ncM)
  ID time  Y  Z
1 705810  1 380 120
2 705810  2 377 120
3 705810  3 460 120
4 705810  4 472 120
5 705810  5 495 120
6 705810  6 566 120
> # ID 2,3,4, Z made missing
> ncM$timeInt = ncM$time - 1
> summary(ncM)
      ID              time              Y              Z              timeInt
Min.   : 705810   Min.   :1.00   Min.   :270.0   Min.   : 64.0   Min.   :0.00
1st Qu.: 847813   1st Qu.:2.75   1st Qu.:395.0   1st Qu.: 97.0   1st Qu.:1.75
Median :1046817   Median :4.50   Median :464.0   Median :106.0   Median :3.50
Mean    :1461655   Mean    :4.50   Mean    :469.9   Mean    :106.1   Mean    :3.50
3rd Qu.:1290819   3rd Qu.:6.25   3rd Qu.:540.0   3rd Qu.:115.0   3rd Qu.:5.25
Max.    :11090821  Max.    :8.00   Max.    :762.0   Max.    :145.0   Max.    :7.00
      NA's      :24
> ncCon2 = lmer(Y ~ Z*timeInt + ( 1 + timeInt | ID), data = ncM) # incl Z in slope L2
> summary(ncCon2)
Linear mixed model fit by REML ['lmerMod']
Formula: Y ~ Z * timeInt + (1 + timeInt | ID)
Data: ncM

REML criterion at convergence: 20256.6

Scaled residuals:
    Min       1Q   Median       3Q      Max
-3.5700 -0.6066 -0.0356  0.5935  3.1695

Random effects:
Groups   Name              Variance Std.Dev. Corr
ID       (Intercept) 194.47    13.945
         timeInt    24.96     4.996    0.38
Residual              401.45    20.036
Number of obs: 2192, groups: ID, 274

Fixed effects:
              Estimate Std. Error t value
(Intercept) 253.32204    8.84339  28.645
Z            0.83725    0.08261  10.135
timeInt      0.82916    2.73189   0.304
Z:timeInt    0.33587    0.02552  13.162

Correlation of Fixed Effects:
          (Intr) Z      timInt
Z         -0.992
timeInt   -0.065  0.065
Z:timeInt  0.065 -0.065 -0.992
> ## full data week 2 Number of obs: 2216, groups: ID, 277, 3 ID's deleted from this analys
>
```

# Package ‘mi’

February 14, 2012

**Version** 0.09-16

**Date** 2012-01-19

**Title** Missing Data Imputation and Model Checking

**Author** Andrew Gelman <gelman@stat.columbia.edu>, Jennifer Hill  
<jh1030@columbia.edu>, Yu-Sung Su <suyusung@tsinghua.edu.cn>, Masanao Ya-  
jima <my2167@columbia.edu>, Maria Grazia Pittau  
<grazia@stat.columbia.edu>

**Maintainer** Yu-Sung Su <suyusung@tsinghua.edu.cn>

**Depends** methods, MASS, nnet, car, arm (>= 1.3-08), Matrix, lme4, R2WinBUGS, abind, car

**Description** Missing-data imputation and model checking

**URL** <http://www.stat.columbia.edu/~gelman/>

**License** GPL (>= 2)

**Repository** CRAN

**Date/Publication** 2012-01-19 20:42:21

## R topics documented:

CHAIN . . . . .	2
convergence.plot . . . . .	3
mi . . . . .	4
mi.binary . . . . .	7
mi.categorical . . . . .	9
mi.completed . . . . .	11
mi.continuous . . . . .	12
mi.count . . . . .	13
mi.fixed . . . . .	15
mi.hist . . . . .	16
mi.info . . . . .	18
mi.info.update . . . . .	19