Statistics 222, Education 351A Autumn 2020

Statistical Methods for Longitudinal Research

Autumn 2020 Remote Asynchronous Instruction

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

To see full course materials from Autumn 2018 go here

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

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

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

Preliminary Course Outline

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

Texts and Resources for Course Content

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

same rate of change. For purposes here use the `sleepstudy` data to fit a mixed-model with all individuals having the same time gradient. Compare to the model in class allowing slopes and levels to differ.

Solution for Review Question 1

2. Orange tree extras. Take the fixed effects from the orange tree nlmer model, "m1" in the class materials, as the parameters of the "average" growth curve for this group of trees. Plot that logistic growth curve (either use a formula for logistic or the `growfit` package has a simple function). Compare the fixed effects from `nlmer` to the results from `nls` for these data. *More challenging* Try to superimpose the group logistic curve (above) onto the plots of the individual tree trajectories (you may want to refer to the plots week1 Aids data).

Solution for Review Question 2

3. Asymptotic regression, SSasymp slide (pdf p.5 of Bates slides, Nonlinear mixed models talk linked in Week 3, Topic 4). Data are from Neter-Wasserman text in file CH13TA04.txt. The outcome variable is manufacturing relative efficiency (RelEff) over 90 weeks duration for two different locations. Plot the RelEff outcome against week for the two locations. Use the SSasymp function for a nlmer fit (or nls if needed) to see whether the asymptote differs for the two locations.

Solution for Review Question 3

4. Quadratic (polynomial) Trends. The book by Mirman resource item 7 *Growth Curve Analysis and Visualization Using R* not surprisingly has some good data examples (primarily psychological learning experiments). Here we use the Chapter 3 data set (sec 3.4) Word Learning. Data at `http://www.danmirman.org/gca/WordLearnEx.txt`. Use the subset TP == Low. How many subjects in that subset? How many observations on each? Accuracy is the outcome measure, the time ordered measure is Block (see Fig 3.7). Investigate a linear trend versus a quadratic trend using mixed effect models.

Solution for Review Question 4

WEEK 3 Exercises

1. Teen age drinking. [note: data location updated 10/12/17]
The UCLA data archive has a comma delimited file (access by
`read.table("https://stats.idre.ucla.edu/stat/r/examples/alda/data/alcohol1_pp.txt", header=T, sep=",")` .
Measurements on 82 adolescents (initial age 14) included 3 time-ordered observations on alcohol use and two background (exogenous) variables: dichotomous `coa` (child of an alcoholic) and measured variable peer (alcohol use by target's peers). Describe the collection of time trajectories in alcohol use. Fit an unconditional mixed model to this collection of time-trajectories and obtain interval estimates for the random and fixed effects. Show a plot for the random effects (subjects) and interpret the fixed effects. Now consider the two exogenous variables. Using conditional models, identify the best fitting model. Interpret the fixed effects for the best fitting model.

2. Vocabulary learning data from test results on file in the Records Office of the Laboratory School of the University of Chicago. Source D R Bock, MSMBR. The data consist of scores, obtained from a cohort of pupils at the eigth through eleventh gade level on alternative forms of the vocabulary section of the Cooperative Reading Tests." There are 64 students in all, 36 male, 28 female (ordered) each with four equally spaced observations (test scores). Wide form of these data are in BOCKwide.dat and I kindly also made a long-form version BOCKlong.dat . Construct the usual collection of individual trajectory displays (either connect-the-dots or compare to a straight-line). Obtain the means (over persons) and plot the group growth curve. Does there appear to be curvature (i.e. deceleration in vocabulary skill growth)?
a. Construct an lmer model with the individual growth curve a quadratic function of grade (year), most convenient to use uncorrelated predictors `grade - mean(grade)` and `(grade - mean(grade))^2`. Fit the lmer model and interpret the fixed and random effects you obtain. Compare the results with a lmer model in which the individual trajectories are straight-line. Use the `anova` model comparison functionality in R (e.g. `anova(modLin, modQuad)` to test whether the quadratic function for individual growth produces a better model fit.
b. Investigate (via lmer model) gender differences (isMale) in vocabulary growth. Fit appropriate lmer models and interpret results,

3. Data on the growth of chicks on different diets. Hand and Crowder (1996), Table A.2, p. 172 Hand, D. and Crowder, M. (1996), Practical Longitudinal Data Analysis, Chapman and Hall, London. The dataset is available as a .R file; easiest to bring this page down to your machine and then `load` into your R-session (or try to load remotely). Here we consider the 20 chicks on Diet 1. (select these). Construct the plots analogous to those for the class example Orange trees: individual chicks frame-by-frame and all chicks on one plot. Fit a nlmer model that allows final weight (asymptote) to differ over chicks (other params fixed). Use ranef (individual estimates) to identify the largest asymptote value and smallest value. Plot the "average" growth curve under diet 1. Compare that nlmer model with a model that does not allow asymptotes to differ. What is your conclusion. Also compare with a nls model that ignores repeated measurements structure (i.e. ignores individual chicks). Compare the average growth curves.

Week 4. Special case of time-1, time-2 data; Traditional measurement of change and more

Lecture Topics

1. Properties of Collections of Growth Curves. class handout
2. Time-1, time-2 data. (paired data)
   The R-package PairedData has some interesting plots and statistical summaries for "before and after" data;
   here is a McNeil plot for Xi.1, Xi.5 in data example
   Paired dichotomous data, McNemar's test (in R, mcnemar.test {stats}), Agresti (2nd ed) sec 10.1
   Also see R-package `PropCIs` Prime Minister example
3. Issues in the Measurement of Change. Class lecture covers Myths 1-6+.
   Slides from Myths talk . Class Handout, Companion for Myths talk
4. Examples for Exogenous Variables and Correlates of Change (use of lagged dependent variables)
   Time-1,time-2 data analysis examples Measurement of change: time-1,time-2 data
   data example for handout scan of regression handout ascii version of data analysis handout
   Extra material for Correlates and predictors of change: time-1,time-2 data
   Rogosa R-session to replicate handout, demonstrate wide-to-long data set conversion, and descriptive fitting of individual growth curves. Some useful plots from Rogosa R-session
   Technical results: Section 3.2.2 esp Equation 27 in Rogosa, D. R., & Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth, Psychometrika, 50, 203-228. Talk slides
5. Comparing groups on time-1, time-2 measurements: repeated measures anova vs lmer OR the t-test

Comparative Analyses of Pretest-Posttest Research Designs, Donna R. Brogan; Michael H. Kutner, *The American Statistician*, Vol. 34, No. 4. (Nov., 1980), pp. 229-232. JSTOR link

urea synthesis, BK data data, long-form
BK plots (by group) BK overview

2017 Analysis handout Extended BK lmer analysis
Additional stuff
BK repeated measures analysis pdf version
Stat141 analysis
archival example analyses. SAS and minitab

Background Readings and Resources
Myths Chapter. Rogosa, D. R. (1995). Myths and methods: "Myths about longitudinal research," plus supplemental questions. In The analysis of change, J. M. Gottman, Ed. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 3-65.
Myths Talk. Rogosa, D. R. (1983)
More stuff (if you don't like the ways I said it)
I noticed John Gottman did a pub rewriting the myths: Journal of Consulting and Clinical Psychology 1993, Vol. 61, No. 6,907-910 The Analysis of Change: Issues, Fallacies, and New Ideas
Also John Willett did a rewrite of the Myths 'cuz I didn't want to reprint it again (or write a new version): Questions and Answers in the Measurement of Change REVIEW OF RESEARCH IN EDUCATION 1988 15: 345
Reliability Coefficients: Background info. Short primer on test reliability Informal exposition in *Shoe Shopping and the Reliability Coefficient*
extensive technical material in Chap 7 Revelle text
A growth curve approach to the measurement of change. Rogosa, David; Brandt, David; Zimowski, Michele Psychological Bulletin. 1982 Nov Vol 92(3) 726-748 APA record direct link
Rogosa, D. R., & Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. Psychometrika, 50, 203-228. available from John Willet's pub page
Demonstrating the Reliability of the Difference Score in the Measurement of Change. David R. Rogosa; John B. Willett Journal of Educational Measurement, Vol. 20, No. 4. (Winter, 1983), pp. 335-343. Jstor
Maris, Eric. (1998). Covariance Adjustment Versus Gain Scores--Revisited. *Psychological Methods*, 3(3) 309-327. apa link
A good R-primer on repeated measures (a lots else). Notes on the use of R for psychology experiments and questionnaires Jonathan Baron, Yuelin Li.
Another version
Multilevel package has behavioral scienes applications including estimates of within-group agreement, and routines using random group resampling (RGR) to detect group effects.
More repeated measures resources: Background primer on analysis of variance (with R); see sections 6.8, 6.9 of *Notes on the use of R for psychology experiments and questionnaires* Jonathan Baron, Yuelin Li. Pdf version The ez package provides extended anova capabities. Examples (blog notes) : Repeated measures ANOVA with R (functions and tutorials) Repeated Measures ANOVA using R Obtaining the same ANOVA results in R as in SPSS - the difficulties with Type II and Type III sums of squares
Application publications, time-1, time-2 Experimental Group Comparisons:
a. Mere Visual Perception of Other People's Disease Symptoms Facilitates a More Aggressive Immune Response *Psychological Science*, April 2010 Pre-post data and difference scores (see Table 1)
b. Guns and testosterone. Guns Up Testosterone, Male Aggression
Guns, Testosterone, and Aggression: An Experimental Test of a Mediational Hypothesis Klinesmith, Jennifer; Kasser, Tim; McAndrew, Francis T, *Psychological Science*. Vol 17(7), Jul 2006, pp. 568-571.

WEEK 4 Review Questions
1. Time1-time2 regressions; Class example
Repeat the handout demonstration regressions using the fallible measures (the X's) from the bottom half of the linked data page. The X's are simply error-in-variable versions of the Xi's: X = Xi + error, with error having mean 0 and variance 10. Compare 5-number summaries for the amount of change from the earliest time "1" to the final observation "5" using the "Xi" measurements (upper frame) and the fallible "X" observations (lower frame).
Solution for Review Question 1
2. (*more challenging*). Use mvrnorm to construct a second artificial data example (n=100) mirroring the week 4 myths data class handout BUT with the correlation between true individual rate of change and W set to .7 instead of 0. Carry out the corresponding regression demonstration.
Solution for Review Question 2
3. *Reliability versus precision demonstration*
Consider a population with true change between time1 and time2 distributed Uniform [99,101] and measurement error Uniform [-1, 1]. If you used discrete Uniform in this construction then you could say measurement of change is accurate to 1 part in a hundred.
Calculate the reliability of the difference score.
Also try error Uniform [-2,2], accuracy one part in 50.
A similar demonstration can be found in my *Shoe Shopping and the Reliability Coefficient*

Solution for Review Question 3
4. Revisit Brogan-Kutner data analysis.
a. Demonstrate the Brogan-Kutner Section 5 equivalences (from paper, shown in class) for repeated measures anova and/or BK lmer analyses.
b. Is amount of gain/decline related to initial status? For the 8 new procedure patients and for the 13 old procedure patients, seperately, estimate the correlation between change and initial status and obtain a confidence interval if possible.
c. Analysis of Covariance. For the Brogan-Kutner data carry out an analysis of covariance (using premeasure as covariate) for the relative effectiveness of the surgery methods. Compare with class analyses.
Slides 203-204 in the Laird-Ware text materials purport to demonstrate that analysis of covariance produces a more precise treatment effect estimate than difference scores (repeated measures anova). What *very* limiting assumption is slipped into their analysis? Can you create a counter-example to their assertion/proof?
Solution for Review Question 4
part c. Solution Notes on the ALA (Laird-Ware) assertion
5. Repeat Brogan-Kutner lmer analyses from lecture. Just another repitition of BK class handout.
Use lmer (or lme) to determine the comparative efficacy of the surgical methods on liver function. Investigate whether a model allowing for pretest differences is helpful.
Solution for Review Question 5

WEEK 4 Exercises

1. Captopril and Blood pressure

The file captopril.dat contains the data shown in Section 2.2 of Verbeke, Introduction to Longitudinal Data Analysis, slides. Captopril is an angiotensin-converting enzyme inhibitor (ACE inhibitor) used for the treatment of hypertension.

a. Smart First Year Student analyses. Use the before and after Spb measurements to examine the improvement (i.e. decrease) in blood pressure. Obtain a five-number summary for observed improvement. What is the correlation between change and initial blood pressure measurement? Obtain a confidence interval for the correlation and show the corresponding scatterplot. What special challenges are present in this analysis?

b. lmer analyses. Try to obtain a good confidence interval for the amount of decline. Obtain a point and interval estimate for the correlation beween initial status and change in Spb.

2. Regression toward the mean? Galton's data on the heights of parents and their children

In the "HistData" or "psych" packages reside the "galton" dataset, the primordial regression toward mean example.

Description: Galton (1886) presented these data in a table, showing a cross-tabulation of 928 adult children born to 205 fathers and mothers, by their height and their mid-parent's height. A data frame with 928 observations on the following 2 variables. parent Mid Parent heights (in inches) child Child Height. Details: Female heights were adjusted by 1.08 to compensate for sex differences. (This was done in the original data set)

Consider "parent" as time1 data and "child" as time2 data and investigate whether these data indicate *regression toward the mean* according to either definition (metric or standardized)? Refer to Section 4 of the Myths chapter supplement (pagination 61-63) for an assessment of regression toward the mean (i.e. counting up number of subjects satisfying regression-toward-mean).

Aside: if you like odd plots, look at the `sunflowerplot` code in the docs for the galton data.

3. Paired and unpaired samples, continuous vs categorical measurements.

Let's use again the 40 subjects in the Review Question 1 "X" data.

a. Measured data. Take the time1 and time5 observations and obtain a 95% Confidence Interval for the amount of change. Compare the width of that interval with a confidence interval for the difference beween the time5 and time1 means if we were told a different group of 40 subjects was measured at each of the time points (data no longer paired).

b. Dichotomous data. Instead look at these data with the criterion that a score of 50 or above is a "PASS" and below that is "FAIL". Carry out McNemar's test for the paired dichotomous data, and obtain a 95% CI for the difference between dependent proportions. Compare that confidence interval with the "unpaired" version (different group of 40 subjects was measured at each of the time points) for independent proportions.

4. *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. Follow the various descriptive and modelling strategies shown in the BK class example.

5. From 2017 *In the news*

The 3 billion dollar (and counting) change score

(items below clipped from 2017 various press reports; we do not have the data)

Sage Therapeutics (NASDAQ:SAGE) surged in response to its announcement of positive results from a Phase 2 clinical trial assessing SAGE-217 for the treatment of adult patients with moderate-to-severe major depressive disorder (MDD), a Fast Track indication in the U.S.[2020 note: name, Zuranolone] It is estimated that there are around 16 million people in the United States with MDD.

SAGE-217, a neuroactive steroid, is next-generation GABA modulator. The GABA system, the major inhibitory signaling pathway in the brain and central nervous system (CNS), plays a key role in regulating CNS function. The company intends to advance the program into Phase 3 development.

The phase 2 looked at the effect of the positive allosteric modulator of the gamma-aminobutyric acid (GABA) receptor as compared to placebo in 89 patients with MDD.

About the Placebo-controlled Phase 2 trial of SAGE-217 in MDD:

In the randomized, double-blind, parallel-group, placebo-controlled trial, 89 eligible patients (with a minimum total score of 22 on the Hamilton Rating Scale for Depression) were stratified based on use of antidepressant treatment (current/stable or not treated/withdrawn >= 30 days) and randomized in a 1:1 ratio to receive SAGE-217 Capsules (30mg) (n=45) or matching placebo (n=44). All doses of study drug were administered at night with food. The study consisted of a 14-day treatment period, and a 4-week follow-up period. The mean HAM-D total scores at baseline were 25.2 for the SAGE-217 group and 25.7 for the placebo group (overall range 22-33), representing patients with moderate to severe MDD. Approximately 90 percent of patients in each group completed the study.

Sage Therapeutics (NASDAQ: SAGE), a clinical-stage biopharmaceutical company developing novel medicines to treat life-altering central nervous system (CNS) disorders, today announced positive top-line results from the Phase 2, double-blind, placebo-controlled clinical trial of SAGE-217 in the treatment of 89 adult patients with moderate to severe major depressive disorder (MDD). In the trial, treatment for 14 days with SAGE-217 was associated with a statistically significant mean reduction in the Hamilton Rating Scale for Depression (HAM-D) 17-Item total score from baseline to Day 15 (the time of the primary endpoint) of 17.6 points for SAGE-217, compared to 10.7 for placebo ($p<0.0001$). Statistically significant improvements were observed in the HAM-D compared to placebo by the morning following the first dose through Week 4 and the effects of SAGE-217 remained numerically greater than placebo through the end of follow-up at Week 6. At Day 15, 64 percent of patients who received SAGE-217 achieved remission, defined as a score of 7 or less on the HAM-D scale, compared with 23 percent of patients who received placebo ($p=0.0005$).

The 89-subject study met its primary endpoint of a statistically significant average reduction in HAM-D score from baseline to day 15 ($p<0.0001$) versus placebo. HAM-D is a rating scale for depression. At day 15, 64% of patients in the treatment group achieved remission compared to 23% for placebo ($p=0.0005$).

There were a total of 89 patients recruited into the study who were either given SAGE-217 or a placebo compound. Patients were treated for a 14 day period and were then measured for clinical outcome using the Hamilton Rating Scale for Depression or HAM-D 17-item total score from baseline. It was shown that SAGE-217 achieved a statistically significant improvement over placebo according to the HAM-D scale. Patients that took SAGE-217 were shown to achieve a 17.6 point improvement at day 15, compared to only a 10.7 point improvement for placebo. That meant that the drug achieved a statistically significant p-value of $p < 0.0001$. It was also noted that 64% of patients who took SAGE-217 achieved MDD remission, compared to only 23% of placebo patients. MDD remission was classified as patients having a HAM-D score of 7 or less. This was the secondary endpoint of the study, which was also achieved.

Investigators saw a statistically significant improvement in SAGE-217 patients on a depression scale the day after the first dose. By the time the two-week treatment period came to an end, the mean score in the SAGE-217 arm had dropped 17.6 points, as compared to a 10.7 point decline in the control group. That seven-point placebo-adjusted improvement was enough for the trial to hit its primary endpoint with a p value of less than 0.0001.

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

```
sagelmer = 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

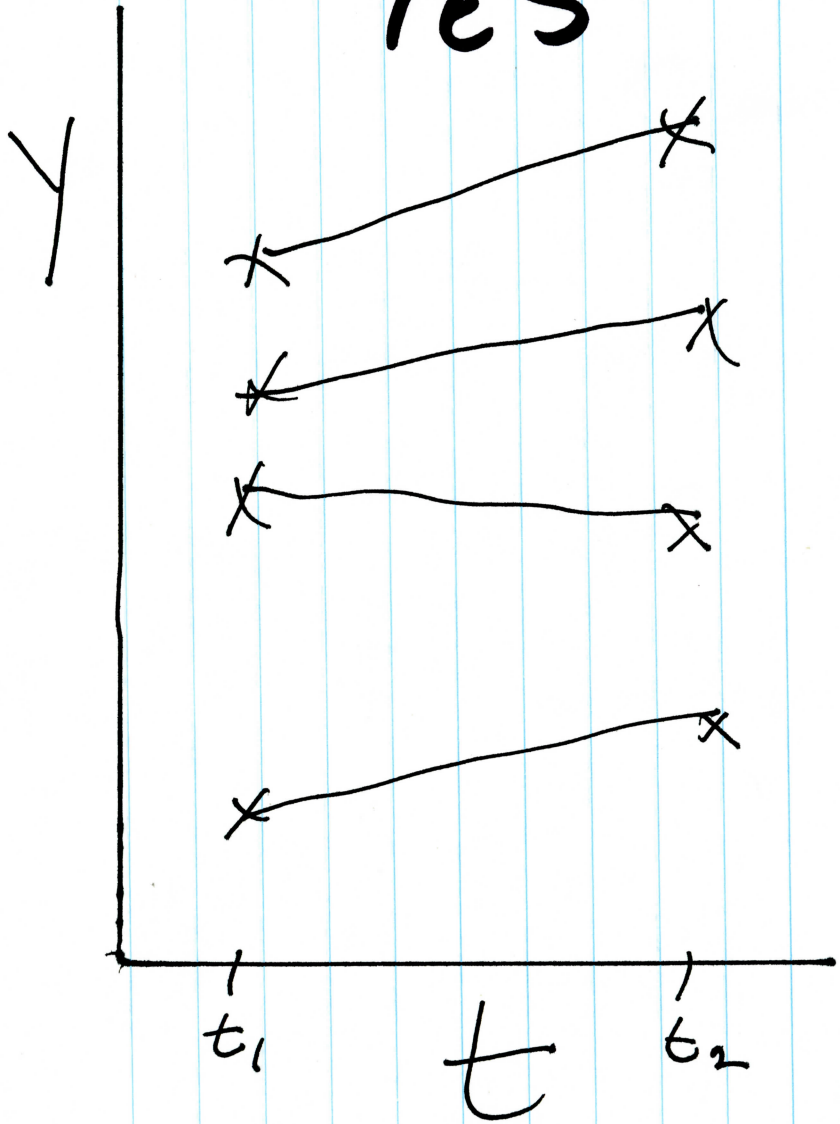
Longitudinal Research Questions

1. Individual and Group Growth
2. Correlates, Predictors of Change Time1-time2 regressions example
3. Stability over Time
4. Comparing Experimental Groups
5. Comparing Nonexperimental Groups
6. Analysis of Reciprocal Effects
7. Growth in Multiple Measures

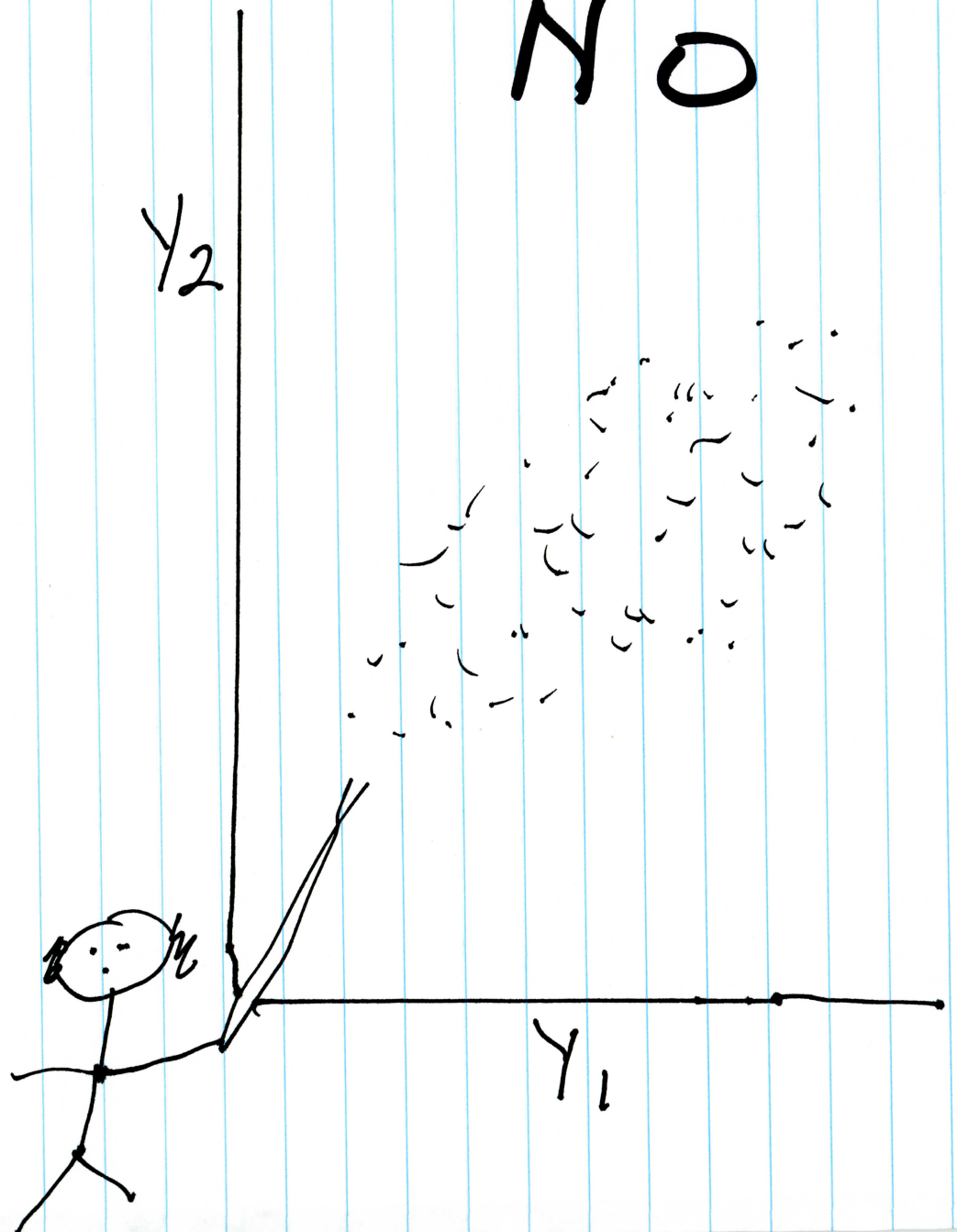
"Different designs and analyses address different sets of these questions."

Time 1, Time 2 data

Yes



No



Properties (Moments of Observables) of Collections of Growth Curves

STAT 222 [redacted] D Rogosa

for indv p $\xi_p(t) = \xi_p(0) + \theta_p t$ t_i $(i = 1, \ldots T)$ p $(p = 1, \ldots n)$

"centering"

$$\xi_p(t) = \xi_p(t^0) + \theta_p (t - t^0)$$

centering, scale

$t^0 = -\sigma_{\xi(0)\theta}/\sigma^2_\theta$

$\rho_{\xi(t^0)\theta} = 0$, min var$(\xi)$

Scale $K = \sigma_{\xi(t^0)}/\sigma_\theta$

time metric

Moments

Covariance $\sigma_{\xi(t_1)\xi(t_2)} = (t_1 - t^0)(t_2 - t^0)\sigma^2_\theta + \sigma^2_{\xi(t^0)}$

Variance $\sigma^2_{\xi(t)} = \sigma^2_{\xi(t^0)} + ((t - t^0)/K)^2 \sigma^2_{\xi(t^0)}$

$$\sigma^2_{\xi(t)}/\sigma^2_{\xi(t^0)} = 1 + \left(\frac{t - t^0}{K}\right)^2$$

Correl change, initial status $\rho_{\theta\xi(t)} = \frac{t - t^0}{[K^2 + (t - t^0)^2]^{1/2}}$

exogenous var w

$$\rho_{w\xi(t)} = \frac{(t - t^0)\rho_{w\theta} + K\rho_{w\xi(t^0)}}{[K^2 + (t - t^0)^2]^{1/2}}$$

where $t^u = t^0 + K\left(\frac{\rho_{w\theta}}{\rho_{w\xi(t^0)}}\right)$ $t^l = t^0 - K\left(\frac{\rho_{w\xi(t^0)}}{\rho_{w\theta}}\right)$

Myths

week 1 example: (p. 64) $\theta \sim U[1, 9]$, $\xi(t^0) \sim U[38, 62]$

$t^0 = 2$ $\sigma^2_\theta = 5.333$ $\sigma^2_{\xi(t^0)} = 48$ $\rho_{w\theta} = 0$ $\rho_{w\xi(t^0)} =$

at time t_i $X_{ip} = \xi_{ip} + \varepsilon$ $\varepsilon \sim (0, \sigma^2_\varepsilon)$ errors in variables

weeks 4 ex $\sigma^2_\varepsilon = 10$

uncorrelated random variables $\xi(t^\circ)$ and θ (e.g., each distribution Gaussian or each distribution Uniform) to generate these parameter values for each p. By doing so, the scale for the time metric $\kappa = \sigma_{\xi(t^\circ)}/\sigma_\theta$ is specified. By then stating the discrete values of the times of observation $\{t_i\} = t_1, \ldots, t_T$, we then have values for the $\xi_p(t_i)$ for $p = 1, \ldots, n$. The exogenous characteristic W is generated with specified mean and variance, specifying the two correlations $\rho_{W\xi(t^\circ)}$ and $\rho_{W\theta}$ (under the constraint $(\rho_{W\xi(t^\circ)})^2 + (\rho_{W\theta})^2 \leq 1$). The final step is to create the fallible observables by the addition of measurement error to the $\xi_p(t_i)$ according to the classical test theory model: $X_p(t_i) = \xi_p(t_i) + \varepsilon_i$ for $p = 1, \ldots, n$.

Consequences for Second Moments

The choices of the values above determine the population values of the familiar second moments of $\xi_p(t_i)$ or $X_p(t_i)$ for the artificial data. In practice, these values of these quantities—variances, correlations, etc.—are often chosen first (say to correspond to values familiar from empirical research or common sense), and then solutions (explicitly or by trial and error) for the corresponding values for the simulation procedure above are obtained. The relations that provide values of these second moments for the $\xi_p(t_i)$ are
variance

$$\sigma^2_{\xi(t)} = \sigma^2_{\xi(t^\circ)} + ((t - t^\circ)/\kappa)^2\sigma^2_{\xi(t^\circ)},$$

covariance (also yields correlation, using above)

$$\sigma_{\xi(t_1)\xi(t_2)} = (t_1 - t^\circ)(t_2 - t^\circ)\sigma^2_\theta + \sigma^2_{\xi(t^\circ)},$$

correlation between change and initial status

$$\rho_{\theta\xi(t)} = \frac{t - t^\circ}{[\kappa^2 + (t - t^\circ)^2]^{1/2}},$$

correlation with exogenous variable, W

$$\rho_{W\xi(t)} = \frac{(t - t^\circ)\rho_{W\theta} + \kappa\rho_{W\xi(t^\circ)}}{[\kappa^2 + (t + t^\circ)^2]^{1/2}}$$

Technical Specifications for Exhibit 1

In terms of the model parameters, the values for the artificial data in Exhibit 1 are $t^\circ = 2$; $\sigma^2_\theta = 5.333$; $\sigma^2_{\xi(t^\circ)} = 48$; for $\theta \sim U[1, 9]$, $\xi(t^\circ) \sim U[38, 62]$. Population mean rate of change is 5, and values of the population correlation coefficients among the $\xi(t_i)$ for observation times $t_1 = 1$, $t_2 = 3$, $t_3 = 5$ are $\rho_{\xi(1)\xi(3)} = .80$, $\rho_{\xi(1)\xi(5)} = .447$, $\rho_{\xi(3)\xi(5)} = .894$. Furthermore, for the fallible measure X with $\text{var}(\varepsilon) = 10$, the population correlations are $\rho_{X(1)X(3)} = .674$, $\rho_{X(1)X(5)} = .391$, $\rho_{X(3)X(5)} = .781$.

ACKNOWLEDGMENTS

I wish to thank Ghassan Ghandour and Haggai Kupermintz for computational and editorial assistance and Gary Williamson for providing the North Carolina data. Programs described in this chapter can be obtained by writing to David Rogosa at rag@leland.stanford.edu.

ADDITIONAL REFERENCES

Alwin, D. F. (1988). Structural equation models in research on human development and aging. In K. W. Schaie, R. T. Campbell, W. M. Meredith, & S. M. Rawlings (Eds.), *Methodological issues in aging research* (pp. 71–170). New York: Springer.

Brogan, D. R., & Kutner, M. H. (1980). Comparative analyses of pretest-postest research designs. *American Statistician, 34*, 229–232.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101*, 147–158.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* New York: Chapman & Hall.

Fern, T. (1975). A Bayesian approach to growth curves. *Biometrika, 62*, 89–100.

Freedman, D. A. (1987). As others see us: A case study in path analysis. *Journal of Educational Statistics, 12*, 101–128.

Freedman, D. A. (1991). Statistical models and shoe leather. In P. Marsden (Ed.), *Sociological Methodology 1991* (pp. 291–313). Washington, DC: American Sociological Association.

Goldstein, H. (1979a). *The design and analysis of longitudinal studies.* London: Academic Press.

Goldstein, H. (1979b). Some models for analysing longitudinal data on educational attainment. *Journal of the Royal Statistical Society A, 142*, 407–442.

Holland, P. W. (1988). Causal inference, path analysis and recursive structural equation models. In C. Clogg (Ed.), *Sociological Methodology 1988* (pp. 449–484). Washington, DC: American Sociological Association.

Hui, S. L., & Berger, J. O. (1983). Empirical Bayes estimation of rates in longitudinal studies. *Journal of the American Statistical Association, 78*, 753–760.

Rogosa, D. R. (1987). Casual models do not support scientific conclusions: A comment in support of Freedman. *Journal of Educational Statistics, 12*, 185–195.

Rogosa, D. R. (1988). Myths about longitudinal research. In K. W. Schaie, R. T. Campbell, W. M. Meredith, and S. C. Rawlings (Eds.), *Methodological issues in aging research* (pp. 171–209). New York: Springer.

Rogosa, D. R. (1991). A longitudinal approach to ATI research: Models for individual growth and models for individual differences in response to intervention. In R. E. Snow and D. E. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach* (pp. 221–248). Hillsdale, NJ: Lawrence Erlbaum.

Rogosa, D. R. (1993). Individual unit models versus structural equations: Growth curve examples. In K. Haagen, D. Bartholomew, and M. Diestler (Eds.), *Statistical modeling and latent variables* (pp. 259–281). Amsterdam: Elsevier North-Holland.

Rogosa, D. R., & Ghandour, G. A. (1989). TIMEPATH: Statistical analysis of individual trajectories. Stanford University.

Package 'PairedData'

July 2, 2014

Type Package

Title Paired Data Analysis

Version 1.0.1

Date 2013-04-18

Author Stephane Champely <champely@univ-lyon1.fr>

Maintainer Stephane Champely <champely@univ-lyon1.fr>

Description This package provides many datasets and a set of graphics (based on ggplot2), statistics, effect sizes and hypothesis tests for analysing paired data with S4 class.

License GPL (>= 2)

Depends methods,graphics,MASS,gld,mvtnorm,lattice,ggplot2

Collate global1.R ClassP1.R

NeedsCompilation no

Repository CRAN

Date/Publication 2013-04-19 07:43:41

R topics documented:

`paired.plotMcNeil` *Parallel lines plot*

Description

Produce a parallel lines plot for paired data.

Usage

```
paired.plotMcNeil(df, condition1, condition2, groups = NULL, subjects,facet = TRUE, ...)
```

Arguments

`df`	a data frame.
`condition1`	name of the variable corresponding to the second sample.
`condition2`	name of the variable corresponding to the first sample.
`groups`	names of the variable corresponding to groups (optional).
`subjects`	names of the variable corresponding to subjects.
`facet`	faceting or grouping strategy for plotting?
`...`	further arguments to be passed to methods.

Value

a graphical object of class ggplot.

Author(s)

Stephane CHAMPELY

References

McNeil, D.R. (1992) On graphing paired data. The American Statistician, 46 :307-310.

See Also

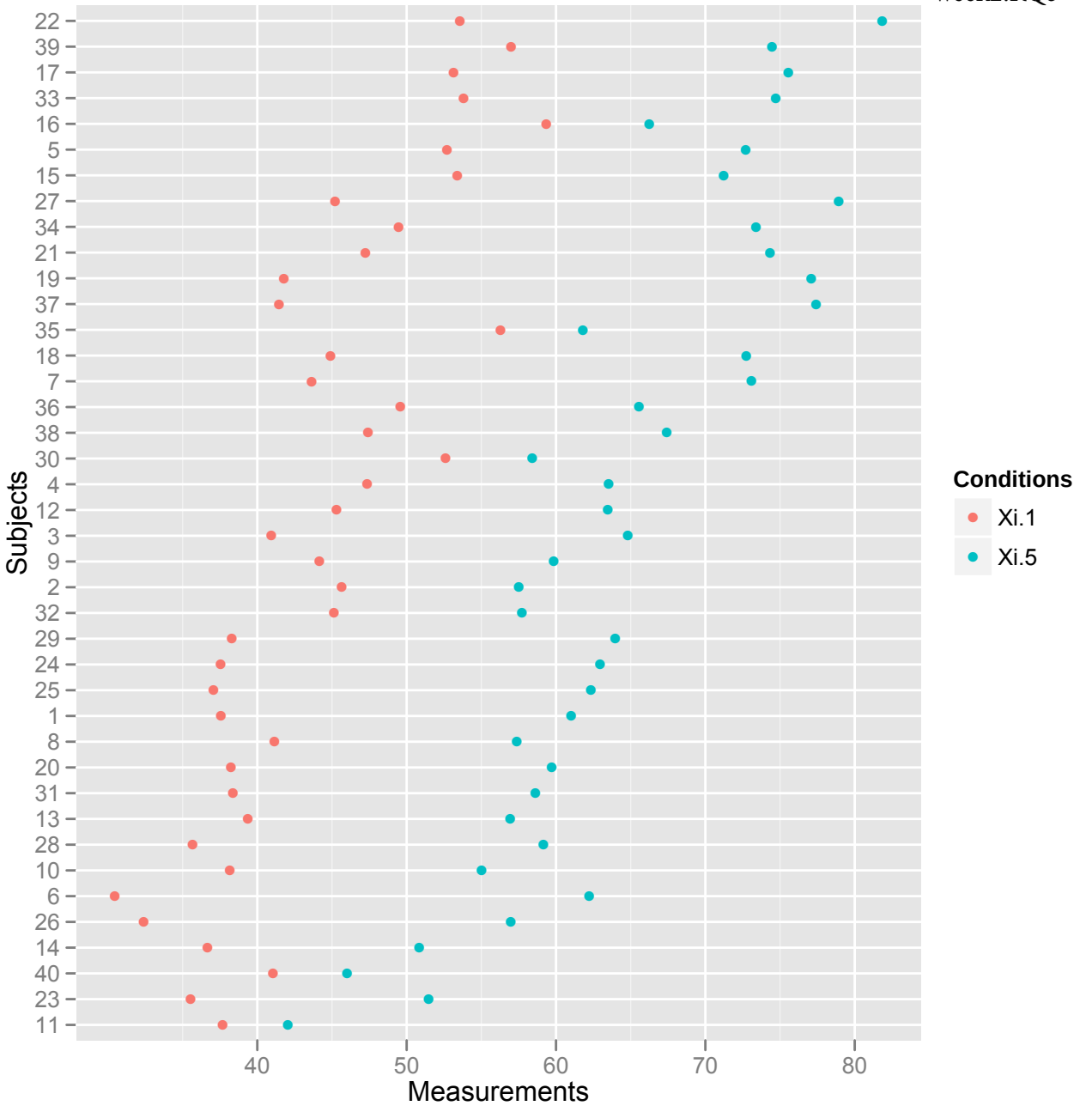
plotBA

Examples

```
data(PrisonStress)
paired.plotMcNeil(PrisonStress,"PSSbefore","PSSafter",subjects="Subject")
```

plotMcNeil(df = week1Xi, condition1 = "Xi.1" ,condition2 = "Xi.5")

week2.RQ3



Chapter 11

1. For a poll of a random sample of 1600 voting-age British citizens, 944 indicated approval of the Prime Minister's performance in office. Six months later, of these same 1600 people, 880 indicated approval. Table 1.22 summarizes results.

Agresti, dichotomous, paired data

Table 1.22:

First Survey	Second Survey: Approve	Second Survey: Disapprove	Total
Approve	794	150	944
Disapprove	86	570	656
Total	880	720	1600

Table 1.23:

Adult Court	Juvenile Court: Rearrest	Juvenile Court: No Rearrest
Rearrest	158	515
No Rearrest	290	1134

Source: Based on a study at the Univ. of Florida by D. Bishop, C. Frazier, L. Lanza-Kaduce, and L. Winner. Thanks to Dr. Larry Winner for showing me these data.

a. Compare the marginal proportions using a confidence interval.

b. Perform McNemar's test, and interpret.

c. Explain why inferences about the difference in approval ratings are more precise than if we had the same sample proportions but with independent samples of size 1600 each.

2. Table 1.23 refers to a sample of juveniles convicted of a felony in Florida in 1987. Matched pairs were formed using criteria such as age and the number of prior offenses. For each pair, one subject was handled in the juvenile court and the other was transferred to the adult court. The response of interest was whether the juvenile was rearrested by the end of 1988. Compare the true proportions rearrested for the adult and juvenile court assignments. Interpret.

3. Table 1.24 shows results when subjects of age between 18 and 29 were asked "Do you think a person has the right to end his or her own life if this person (1) has an incurable disease? (2) is tired of living an ready to die?"

a. Compare the marginal proportions using a confidence interval.

b. Perform McNemar's test, and interpret.

Table 1.24:

Suicide	Let Patient Die: Yes	Let Patient Die: No	Total
Yes	1097	90	1187
No	203	435	638

Source: 1994 General Social Survey

 study descr starts SW p. 441 sec 10.8

older siblings and in 20 of the 114 younger siblings.[59] These data are shown Table 10.25.

TABLE 10.25 HIV Infection Data

		Older Sibling	Younger Sibling
HIV?	Yes	19	20
	No	95	94
	Total	114	114

At first glance, it might appear that a regular chi-square test could be used to te the null hypothesis that the probability of HIV infection is the same for older si lings as for younger siblings. However, as we stated in Section 10.6, for the ch square test to be valid the two samples—of 114 older siblings and of 114 young siblings—must be independent of each other. In this case the samples are clear dependent. Indeed, these are paired data, with a family generating the pair (old sibling, younger sibling).

Table 10.26 presents the data in a different format. This format helps foc attention on the relevant part of the data.

TABLE 10.26 HIV Infection Data Shown by Pairs

		Younger Sibling HIV?	
		Yes	*No*
Older sibling HIV?	Yes	2	17
	No	18	77

From Table 10.26 we can see that there are 79 pairs in which both siblin have the same HIV status: 2 are "yes/yes" pairs and 77 are "no/no" pairs. These pairs, which are called **concordant pairs**, do not help us determine whether HI infection is more likely for younger siblings than for older siblings. The remaini 35 pairs—17 "yes/no" pairs and 18 "no/yes" pairs—do provide information on t relative likelihood of HIV infection for older and younger siblings. These pairs a called **discordant pairs**; we will focus on these 35 pairs in our analysis.

If the chance of HIV infection is the same for older siblings as it is f younger siblings, then the two kinds of pairs—"yes/no" and "no/yes"—are equal likely. Thus, the null hypothesis

H_0: the probability of HIV infection is the same for older siblings as it is f younger siblings

is equivalent to

$$H_0: \text{among discordant pairs, } \Pr(\text{"yes/no"}) = \Pr(\text{no/yes}) = \frac{1}{2}$$

McNemar's Test

The hypothesis that discordant pairs are equally likely to be "yes/no" or "no/ye can be tested with the chi-square goodness-of-fit test developed in Section 10.2. T application of the goodness-of-fit test is known as **McNemar's test** and has

TABLE 10.27 A General Table of Paired Proportion Data

	Yes	No
Yes	n_{11}	n_{12}
No	n_{21}	n_{22}

articularly simple form.* Let n_{11} denote the number of "yes/yes" pairs, n_{12} the umber of "yes/no" pairs, n_{21} the number of "no/yes" pairs, and n_{22} the number of no/no" pairs, as shown in Table 10.27. If H_0 is true, the expected number of yes/no" pairs is $\frac{n_{12} + n_{21}}{2}$, as is the expected number of "no/yes" pairs. Thus, the est statistic is

$$\chi_s^2 = \frac{\left(n_{12} - \frac{(n_{12} + n_{21})}{2}\right)^2}{\frac{(n_{12} + n_{21})}{2}} + \frac{\left(n_{21} - \frac{(n_{12} + n_{21})}{2}\right)^2}{\frac{(n_{12} + n_{21})}{2}}$$

which simplies to

$$\chi_s^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

The distribution of χ_s^2 under the null hypothesis is approximately a χ^2 distribution with 1 degree of freedom.

HIV Transmission to Children. For the data given in Example 10.38, $n_{12} = 17$ and $n_{21} = 18$. Thus,

$$\chi_s^2 = \frac{(17 - 18)^2}{17 + 18} = 0.0286$$

From Table 9 we see that the P-value is greater than .20. (Using a computer gives $P = .87$.) The data are very much consistent with the null hypothesis that the probability of HIV infection is the same for older siblings as it is for younger siblings. ■

Exam

Exercises 10.63–10.65

10.63 As part of a study of risk factors for stroke, 155 women who had experienced a hemorrhagic stroke (cases) were interviewed. For each case, a control was chosen who had not experienced a stroke; the control was matched to the case by neighborhood of residence, age, and race. Each woman was asked whether she used oral

* The null hypothesis tested by McNemar's test can also be tested by using the binomial distribution. The null hypothesis states that among discordant pairs, Pr("yes/no") = Pr("no/yes") = $\frac{1}{2}$. Thus, under the null hypothesis, the number of "yes/no" pairs has a binomial distribution with n = the number of discordant pairs and $p = .5$.

Package 'PropCIs'

February 23, 2018

Type Package

Title Various Confidence Interval Methods for Proportions

Version 0.3-0

Date 2018-02-22

Author Ralph Scherer

Maintainer Ralph Scherer <shearer.ra76@gmail.com>

Description
Computes two-sample confidence intervals for single, paired and independent proportions.

License GPL

URL https://github.com/shearer/PropCIs

BugReports https://github.com/shearer/PropCIs/issues

LazyLoad yes

NeedsCompilation no

Repository CRAN

Date/Publication 2018-02-23 16:49:49 UTC

R topics documented:

`diffpropci.mp` *Adjusted Wald interval for a difference of proportions with matched pairs*

Description

Adjusted Wald interval for a difference of proportions with matched pairs. This is the interval called Wald+2 in Agresti and Min (2005). Adds 0.5 to each cell before constructing the Wald CI

Usage

```
diffpropci.mp(b, c, n, conf.level)
```

Arguments

`b`	off-diag count
`c`	off-diag count
`n`	sample size
`conf.level`	confidence coefficient $1 - \alpha$

Details

The interval is truncated, when it overshoots the boundary

Value

A list with class '"htest"' containing the following components:

`conf.int`	a confidence interval for the difference in proportions.
`estimate`	estimated difference in proportions

References

Agresti, A. and Min, Y. (2005) Simple improved confidence intervals for comparing matched proportions. *Statistics in Medicine* 24 (5), 729–740.

Examples

```
diffpropci.mp(b = 40, c = 20, n = 160, conf.level = 0.95)
```

McNemar (via Agresti) Time1-Time2 Dichotomous data

```
ratings <- matrix(c(794,150, 86, 570), ncol=2, byrow=TRUE,
+ dimnames = list("First Survey" = c("Approve", "Disapprove"),
+ "Second Survey" = c("Approve", "Disapprove")))
> mcnemar.test(ratings, correct=FALSE)

-----------------------------
R-session
> ?mcnemar.test
> ratings <- matrix(c(794,150, 86, 570), ncol=2, byrow=TRUE,
+  dimnames = list("First Survey" = c("Approve", "Disapprove"),
+  "Second Survey" = c("Approve", "Disapprove")))
>  mcnemar.test(ratings, correct=FALSE)

        McNemar's Chi-squared test

data:  ratings
McNemar's chi-squared = 17.3559, df = 1, p-value = 3.099e-05

> ratings
            Second Survey
First Survey Approve Disapprove
  Approve        794         150
  Disapprove      86         570
> sqrt(17.36)
[1] 4.166533
> #Agresti p.411; decline in approval from .59 to .55 (signif) CI (-.06,-.02)
# see R-package "PropCIs"

> install.packages("PropCIs")
Installing package(s) into 'C:/Users/rag/Documents/R/win-library/2.14'
(as 'lib' is unspecified)
--- Please select a CRAN mirror for use in this session ---
trying URL 'http://cran.stat.ucla.edu/bin/windows/contrib/2.14/PropCIs_0.1-7.zip'
Content type 'application/zip' length 48541 bytes (47 Kb)
opened URL
downloaded 47 Kb
package 'PropCIs' successfully unpacked and MD5 sums checked
The downloaded packages are in
        C:\Users\rag\AppData\Local\Temp\RtmpINgSzT\downloaded_packages

> library(PropCIs)
Warning message:
package 'PropCIs' was built under R version 2.14.2

> diffpropci.mp(150,86, 1600, .95)

data:

95 percent confidence interval:
 -0.05868294 -0.02121719
sample estimates:
[1] -0.03995006

References
Agresti, A. and Min, Y. (2005) Simple improved confidence intervals
for comparing matched proportions. Statistics in Medicine 24 (5),
729–740.
```

Example from exercises

2.2 Captopril Data

- Taken from Hand, Daly, Lunn, McConway, & Ostrowski (1994)
- 15 patients with hypertension
- The response of interest is the supine blood pressure, before and after treatment with CAPTOPRIL

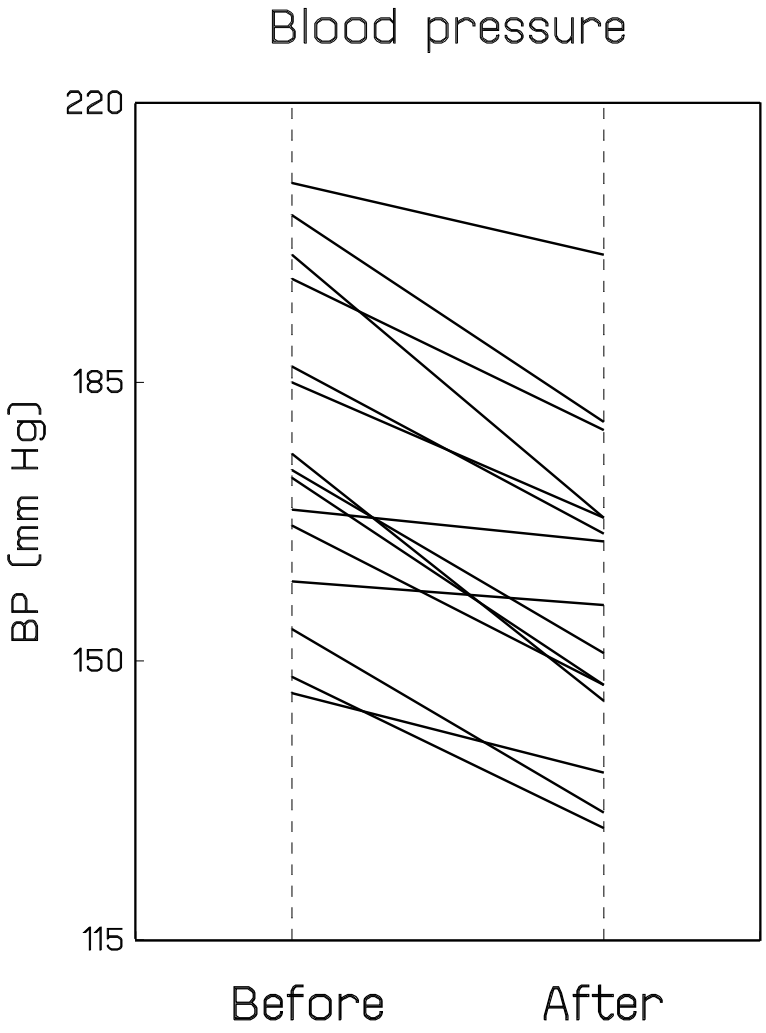
	Before		After	
Patiënt	SBP	DBP	SBP	DBP
1	210	130	201	125
2	169	122	165	121
3	187	124	166	121
4	160	104	157	106
5	167	112	147	101
6	176	101	145	85
7	185	121	168	98
8	206	124	180	105
9	173	115	147	103
10	146	102	136	98
11	174	98	151	90
12	201	119	168	98
13	198	106	179	110
14	148	107	129	103
15	154	100	131	82

Fitting a line to two points

- Research question:

  How does treatment affect BP ?

- Remarks:
  - ▷ Paired observations:
    Most simple example of longitudinal data
  - ▷ Much variability between subjects



measurement of change, this paper does chart a very different direction from that seen in the behavioral sciences literature over the last 50 years. The intended impact of this paper is to direct the emphasis in the measurement of change to the statistical analysis of collections of individual time paths.

The best example of the proper approach to the study of change is the use of models for individual growth in Bock (1976). Also, a kindred perspective in modeling individual growth is seen in the work of Weisberg and Bryk on the estimation of treatment effects from nonequivalent group designs (Bryk & Weisberg, 1977; Bryk, Strenio, & Weisberg, 1980; Weisberg, 1979). The antithesis of our approach is represented by attempts to analyze "change" through covariance structure models for relations among variables as in Sörbom (1976) or similarly, through simpler regression models as in the texts by Cohen and Cohen (1975, chap. 9), Goldstein (1980, chap. 5), and Kessler and Greenberg (1981).

The body of this paper is composed of results and observations that follow naturally from the models for individual growth. These results are used for two purposes. First, much of the detailed discussion of this paper is devoted to clearing up misconceptions and resolving extant confusions in the psychometric work on the measurement of change. Second, the framework introduced for the measurement of change is designed to encourage further methodological work and to improve empirical investigations of change, with an emphasis on the use of multiwave data.

Although this paper strives to be comprehensive, many relevant topics in the study of individual change could not be included. Among these topics are models for change in binary variables (Plewis, 1981), the construction of test items and tests for use in the measurement of change (Saupe, 1966), and the scaling of test-item data using Item Response Theory methods (see Bock, 1976). Also, efficient design for the estimation of individual growth curves, that is, determination of the number and spacing of observations, is an important omission. Finally, we remind the reader that, except for occasional comment, we do not address other purposes for the analysis of longitudinal data, such as correlates of change, comparison of change across experimental or nonequivalent groups, or the study of reciprocal effects.

The major messages of this investigation are summarized in the following series of mottos.

Mottos for the Measurement of Individual Change

1. Individual time paths are the proper focus for the analysis of change.
2. A model for individual change is useful for the measurement of change.
3. The collection of individual X on t regression functions is the key initial summary of the data. The X_2 on X_1 regression is not a good source of information on individual change.
4. Two waves of data are better than one, but maybe not much better. Two data points provide meager information on individual change, and thus the measurement of change often will require more than the traditional pre–post data.
5. When only two waves of data are available, the difference score is a natural and useful estimate of individual change.
6. There's more than one way to judge a measure of change. Reliability is not the "be all and end all" in the measurement of change. Statistical properties are important.
   a. Low reliability does not necessarily mean lack of precision.
   b. The difference between two fallible measures can be nearly as reliable as the measures themselves.
7. The correlation between true change and true initial status (zero or otherwise) is an interesting fact of life. Use of fallible scores to construct poor estimates of this correlation does not invalidate the difference score as a measure of individual change.
8. Measures of individual change can be "improved" by incorporating information from all n persons into the measure of change.
9. The residual change question—How much would person j have changed if everyone had started out equal?—is extremely difficult to answer and is logi-

cally subordinate to the question—What is the (true) change of person j? First things first in the measurement of change.

10. When used wisely, multiwave data will yield far better determinations of individual change than will two-wave data.

References

Bereiter, C. Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.), *Problems in the measurement of change.* Madison, Wisc.: University of Wisconsin Press, 1963.

Blomqvist, N. On the relation between change and initial value. *Journal of the American Statistical Association,* 1977, *72,* 746–749.

Bock, R. D. Basic issues in the measurement of change. In D. N. M. de Gruijter & L. J. T. van der Kamp (Eds.) *Advances in psychological and educational measurement.* New York: Wiley, 1976.

Bohrnstedt, G. W. Observations on the measurement of change. In E. F. Borgatta (Ed.), *Sociological Methodology 1969.* San Francisco: Jossey-Bass, 1969.

Bond, L. On the base-free measure of change proposed by Tucker, Damarin, and Messick. *Psychometrika,* 1979, *44,* 351–355.

Bryk, A. S., & Weisberg, H. I. Use of the nonequivalent control group design when subjects are growing. *Psychological Bulletin,* 1977, *84,* 950–962.

Bryk, A. S., Strenio, J. F., & Weisberg, H. I. A method for estimating treatment effects when individuals are growing. *Journal of Educational Statistics,* 1980, *5,* 5–34.

Cohen, J., & Cohen, P. *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, N.J.: Erlbaum, 1975.

Cox, D. R., & Hinkley, D. V. *Theoretical statistics.* London: Chapman and Hall, 1974.

Cronbach, L. J., & Furby, L. How should we measure "change"—or should we? *Psychological Bulletin,* 1970, *74,* 68–80.

DuBois, P. H. *Multivariate correlational analysis.* New York: Harper, 1957.

Efron, B., & Morris, C. Limiting the risk of Bayes and empirical Bayes estimators-Part II: The empirical Bayes case. *Journal of the American Statistical Association,* 1972, *67,* 130–139.

Efron, B., & Morris, C. Stein's paradox in statistics. *Scientific American,* 1977, *236*(5), 119–127.

Fearn, T. A Bayesian approach to growth curves. *Biometrika,* 1975, *62,* 89–100.

Furby, L. Interpreting regression toward the mean in developmental research. *Developmental Psychology,* 1973, *8,* 172–179.

Glass, G. V. Response to Traub's "Note on the reliability of residual change scores." *Journal of Educational Measurement,* 1968, *5,* 265–267.

Goldstein, H. *The design and analysis of longitudinal studies.* London: Academic Press, 1979.

Guire, K. E., & Kowalski, C. J. Mathematical description and representation of developmental change functions on the intra- and interindividual levels. In J. R. Nesselroade & P. B. Baltes (Eds.), *Longitudinal research in the study of behavior and development.* New York: Academic Press, 1979.

Hannan, M. T., & Tuma, N. B. Methods for temporal analysis. In A. Inkeles (Ed.), *Annual review of sociology* (Vol. 5). Palo Alto, Calif.: Annual Reviews, 1979.

Kelley, T. L. *Interpretation of educational measurements.* Yonkers-on-Hudson, N.Y.: World Book, 1927.

Kessler, R. C. The use of change scores as criteria in longitudinal survey research. *Quality and Quantity,* 1977, *11,* 43–66.

Kessler, R. C., & Greenberg, D. F. *Linear panel analysis.* New York: Academic Press, 1981.

Knapp, T. R. The (un)reliability of change scores in counseling research. *Measurement and Evaluation in Guidance,* 1980, *13,* 149–157.

Labouvie, E. W. Measurement of individual differences in intraindividual change. *Psychological Bulletin,* 1980, *88,* 54–59.

Lindley, D. V., & Smith, A. F. M. Bayes estimates for the linear model. *Journal of the Royal Statistical Society* (Series B), 1972, *34,* 1–41.

Linn, R. L., & Slinde, J. A. The determination of the significance of change between pre- and posttesting periods. *Review of Educational Research,* 1977, *47,* 121–150.

Lord, F. M. The measurement of growth. *Educational and Psychological Measurement,* 1956, *16,* 421–437.

Lord, F. M. Further problems in the measurement of growth. *Educational and Psychological Measurement,* 1958, *18,* 437–454.

Lord, F. M. Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change.* Madison, Wisc.: University of Wisconsin Press, 1963.

Lord, F. M. Significance test for a partial correlation corrected for attenuation. *Educational and Psychological Measurement,* 1974, *34,* 211–220.

Lord, F. M., & Novick, M. N. *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley Publishing Company, 1968.

Manning, W. H., & DuBois, P. H. Correlational methods in research on human learning. *Perceptual and Motor Skills,* 1962, *15,* 287–321.

Marks, E., & Martin, C. G. Further comments relating to the measurement of change. *American Educational Research Journal,* 1973, *10,* 179–191.

Markus, G. B. *Analyzing panel data.* Beverly Hills, Calif.: Sage Publications, 1979.

McNemar, Q. On growth measurement. *Educational and Psychological Measurement,* 1958, *18,* 47–55.

Messick, S. Denoting the base-free measure of change. *Psychometrika,* 1981, *46,* 215–217.

Mosier, C. I. On the reliability of a weighted composite. *Psychometrika,* 1943, *8,* 161–168.

Nesselroade, J. R., Stigler, S. M., & Baltes, P. B. Regression toward the mean and the study of change. *Psychological Bulletin,* 1980, *88,* 622–637.

O'Connor, E. F. Extending classical test theory to the measurement of change. *Review of Educational Research,* 1972, *42,* 73–98.

Plewis, I. A comparison of approaches to the analysis of longitudinal categoric data. *British Journal of Mathematical and Statistical Psychology,* 1981, *34,* 118–123.

Myths

about

LONGITUDINAL RESEARCH

DAVID ROGOSA

STANFORD UNIV.

LONGITUDINAL PANEL DATA

OBSERVATIONS X_{ip}

TAKEN AT TIME t_i $(i = 1, \ldots T)$

FOR INDIVIDUAL p $(p = 1, \ldots, n)$

T "WAVES" OF DATA

MEASUREMENT MODEL

$$X_{ip} = \xi_{ip} + \epsilon_{ip}$$

"TRUE SCORE" ξ_{ip}

reliability coeff
Var(Xi)/Var(X)

GROWTH MODELS

$$\xi_p(t) = f(\xi, t)$$

Collection of Growth Curves

For individual p, growth curve for single measure $\xi_p(t)$

Parameters of growth curve vary over p

Examples: Straight-line growth

$$\xi_p(t) = \xi_p(0) + \theta_p t$$

Exponential growth

$$\xi_p(t) = \lambda_p - (\lambda_p - \xi_p(0))e^{-\gamma_p t}$$

Alternative models

Autoregressive process/

Simplex models

Systematic Individual Differences in Growth

Stage 1 Growth curve

$$\xi_p(t) = \xi_p(0) + \theta_p t$$

Stage 2 Parameters of growth curve depend on individual characteristics $\underline{w}_p$ (vector, scalar)

e.g. $E(\theta_p | \underline{w}_p) = \mu_\theta + \gamma(\underline{w}_p - \mu_w)$

can also model level w/ same or diff w.

Myths about Longitudinal Research

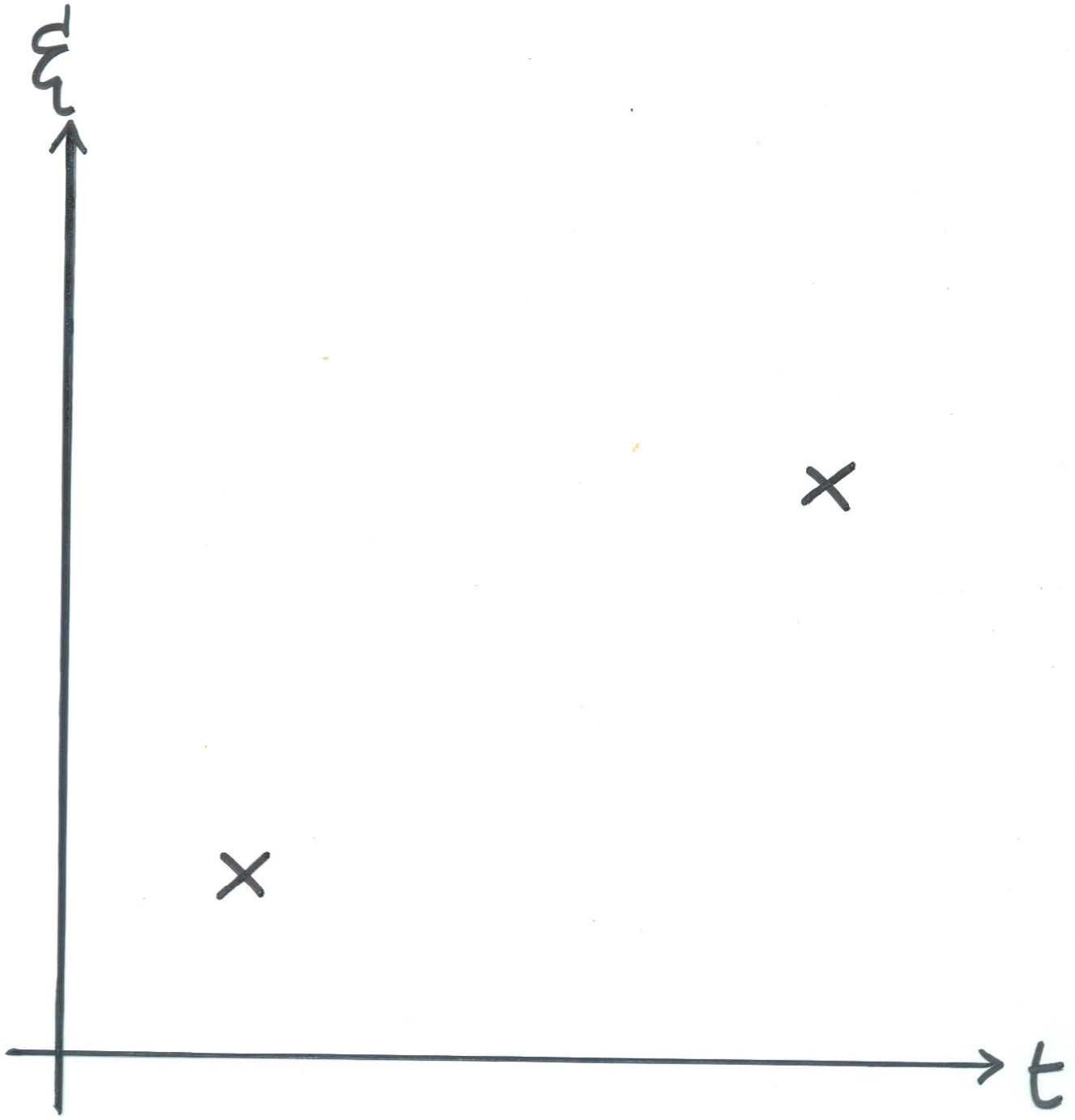
1. Two Observations a longitudinal study make.
2. The difference score is intrinsically unreliable and unfair
3. You can determine from the correlation matrix for the longitudinal data whether or not you are measuring the same thing over time
4. The correlation between change and initial status is:
   (a) negative; (b) zero; (c) positive; (d) all of the above
5. You can't avoid regression toward the mean
6. Residual change cures what ails the difference score
7. Analyses of covariance matrices inform about change
8. Stability coefficients estimate:
   (a) the consistency over time of an individual; (b) the consistency over time of an average individual; (c) the consistency over time of individual differences; (d) none of the above; (e) some of the above
9. Casual analyses support causal inferences about reciprocal effects

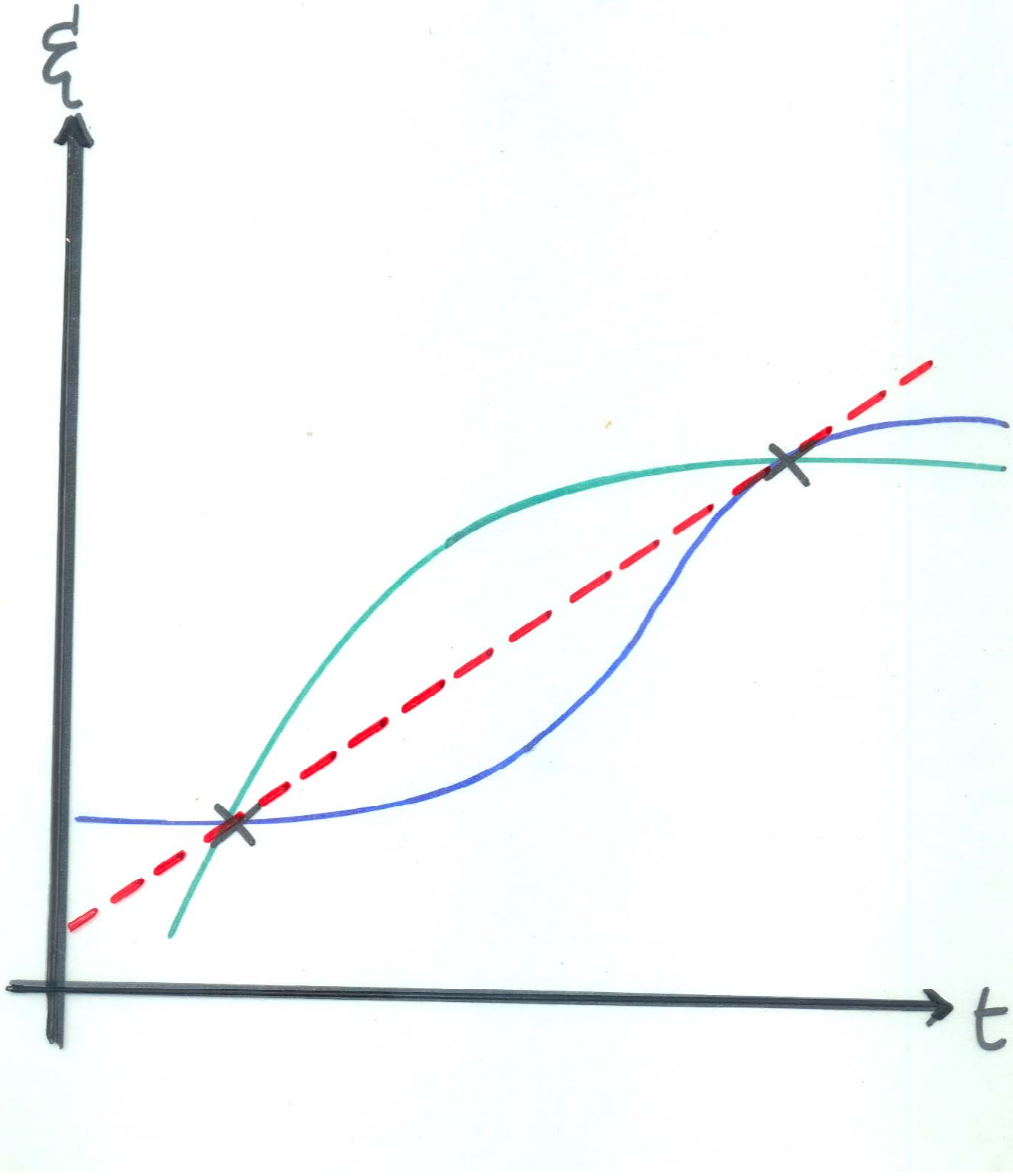
META-MYTH

"Investigators who ask questions regarding gain scores would ordinarily be better advised to frame their questions in other ways."

(Cronbach & Furby, 1970)

TWO
OBSERVATIONS
A
LONGITUDINAL
STUDY
MAKE


ξ
t


ξ
t

Amount Of Change

Individual Growth Curve

$$\xi_p(t) = \lambda_p - [\lambda_p - \xi_p(0)]\, e^{-\gamma_p t}$$

Amount of Change between t and $t+\tau$

$$\Delta_p(t, t+\tau) = [\lambda_p - \xi_p(0)][1 - e^{-\gamma_p \tau}]\, e^{-\gamma_p t}$$

Example – six growth curves.

Message: Amount of change no guide to individual differences in growth.

Myths Companion

STAT 222
week 2
D Rogosa

M1 individual growth $\xi_p(t) = f(\xi, t)$
proportional growth to asymptote λ_p
Amount of Change $\xi_p(t) = \lambda_p - (\lambda_p - \xi_p(0))\, e^{-\gamma_p t}$
$A_p(t, t+c) = [\lambda_p - \xi_p(0)][1 - e^{-\gamma_p c}]\, e^{-\gamma_p(t)}$ trading places

M2 $X_{ip} = \xi_{ip} + \varepsilon$ (fallible score, see back)
$D = X_2 - X_1$, $E(D) = \xi_2 - \xi_1$
$\sigma_\Delta^2 = (t_2 - t_1)^2 \sigma_\Theta^2$ individual differences in change
reliability $\rho(D) = \dfrac{\sigma_\Delta^2}{\sigma_\Delta^2 + \sigma_{\varepsilon_2 - \varepsilon_1}^2}$

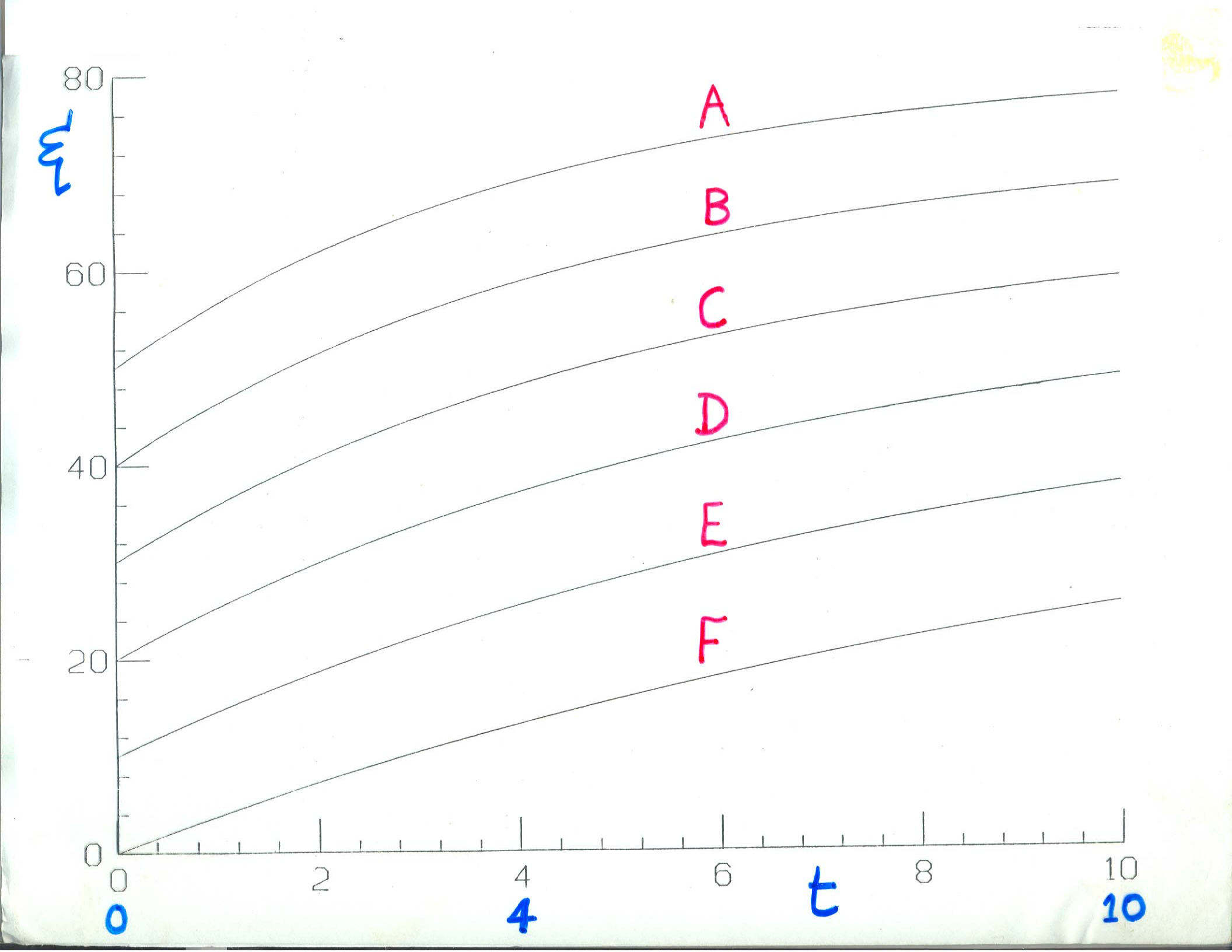
M3 $\rho_{\xi(t)\,\xi(t+c)}$ on back, depends on choice of t, $t+c$
for $\rho_{\xi_1 \Delta} = 0$ $\rho_{\xi_1 \xi_2} = (1 + \sigma_\Delta^2 / \sigma_{\xi_1}^2)^{-1/2}$

M4 $\rho_{\xi(t_1)\Theta}$ on back depends on $t_1 - t^0$
Bias of $r_{X_1 D}$ $E(r_{X_1 D}) = \rho_{\xi(t_1)\Theta} \sqrt{\rho(X_1)\rho(D)} - \dfrac{\sigma_{\varepsilon_1}^2 - \sigma_{\varepsilon_1 \varepsilon_2}}{\sigma_{X_1} \sigma_D}$
proportional bias; additional downward

M5 Standardized Tautology
$\dfrac{E(\xi_2 \mid \xi_1 = c) - \mu_{\xi_2}}{\sigma_{\xi_2}} < \dfrac{c - \mu_{\xi_1}}{\sigma_{\xi_1}} \Rightarrow \rho_{\xi_1 \xi_2} < 1$

In metric of data $E(\xi_2 \mid \xi_1 = c) - \mu_{\xi_2} < c - \mu_{\xi_1}$
$\Rightarrow \rho_{\xi_1 \Delta} < 0$

M6/7 resid change in sample $X_2 - X_1$ bias poor reliability
Correlation $\rho[\Delta \cdot \xi(t_1)]\, \omega$ vs $\rho\, \omega$
$\Delta \cdot \xi(t_1) = \xi(t_2) \cdot \xi(t_1)$


80
60
40
20
0
ξ
A
B
C
D
E
F
0
2
4
6
8
10
t
0
4
10

Amount of True Change

$$\Delta(t_I, t_I + 1)$$

	$t_I = 0$	$t_I = 4$	$t_I = 10$
A	6.64	2.44	.54
B	6.32	2.62	.70
C	5.88	2.75	.88
D	5.37	2.81	1.07
E	4.63	2.75	1.26
F	3.81	2.55	1.40

The
difference score
is
intrinsically
unreliable
and
unfair

Myths Companion

STAT 222
week 2
D Rogosa

M1 individual growth $\xi_p(t) = f(\xi, t)$
proportional growth to asymptote λ_p
Amount of Change $\xi_p(t) = \lambda_p - (\lambda_p - \xi_p(0))\, e^{-\gamma_p t}$
$A_p(t, t+c) = [\lambda_p - \xi_p(0)][1 - e^{-\gamma_p c}]\, e^{-\gamma_p(t)}$ trading places

M2 $X_{ip} = \xi_{ip} + \varepsilon$ (fallible score, see back)
$D = X_2 - X_1$, $E(D) = \xi_2 - \xi_1$
$\sigma^2_\Delta = (t_2 - t_1)^2 \sigma^2_\omega$ individual differences in change
reliability $\rho(D) = \dfrac{\sigma^2_\Delta}{\sigma^2_\Delta + \sigma^2_{\varepsilon_2} - \varepsilon_1}$

M3 $\rho_{\xi(t)\,\xi(t+c)}$ on back, depends on choice of t, $t+c$
for $\rho_{\xi,\Delta} = 0$ $\rho_{\xi_1 \xi_2} = (1 + \sigma^2_\Delta / \sigma^2_{\xi_1})^{-1/2}$

M4 $\rho_{\xi(t_1)\omega}$ on back depends on $t_1 - t^0$
Bias of $r_{X,D}$ $E(r_{X,D}) = \rho_{\xi(t_1)\omega} \sqrt{\rho(X_1)\rho(D)} - \dfrac{\sigma^2_{\varepsilon_1} - \sigma_{\varepsilon_1 \varepsilon_2}}{\sigma_{X_1}\sigma_D}$
proportional bias; additional downward

M5 Standardized Tautology
$\dfrac{E(\xi_2 \mid \xi_1 = c) - \mu_{\xi_2}}{\sigma_{\xi_2}} < \dfrac{c - \mu_{\xi_1}}{\sigma_{\xi_1}} \Rightarrow \rho_{\xi_1 \xi_2} < 1$

In metric of data $E(\xi_2 \mid \xi_1 = c) - \mu_{\xi_2} < c - \mu_{\xi_1}$
$\Rightarrow \rho_{\xi_1 \Delta} < 0$

M6/7 resid change in sample $X_2 - X_1$ biased poor reliability
Correlation $\rho[\Delta \cdot \xi(t_1)]\omega$ vs $\rho\omega$
$\Delta \cdot \xi(t_1) = \xi(t_2) \cdot \xi(t_1)$

Traditional Tabulation of $\rho(D)$ (from Linn & Slinde, 1977)

	$\rho(X)$		
$\rho_{X_1X_2}$	.7	.8	.9
.5	.40	.60	.80
.6	.25	.50	.75
.7	.00	.33	.67
.8	—	.00	.50
.9	—	—	.00

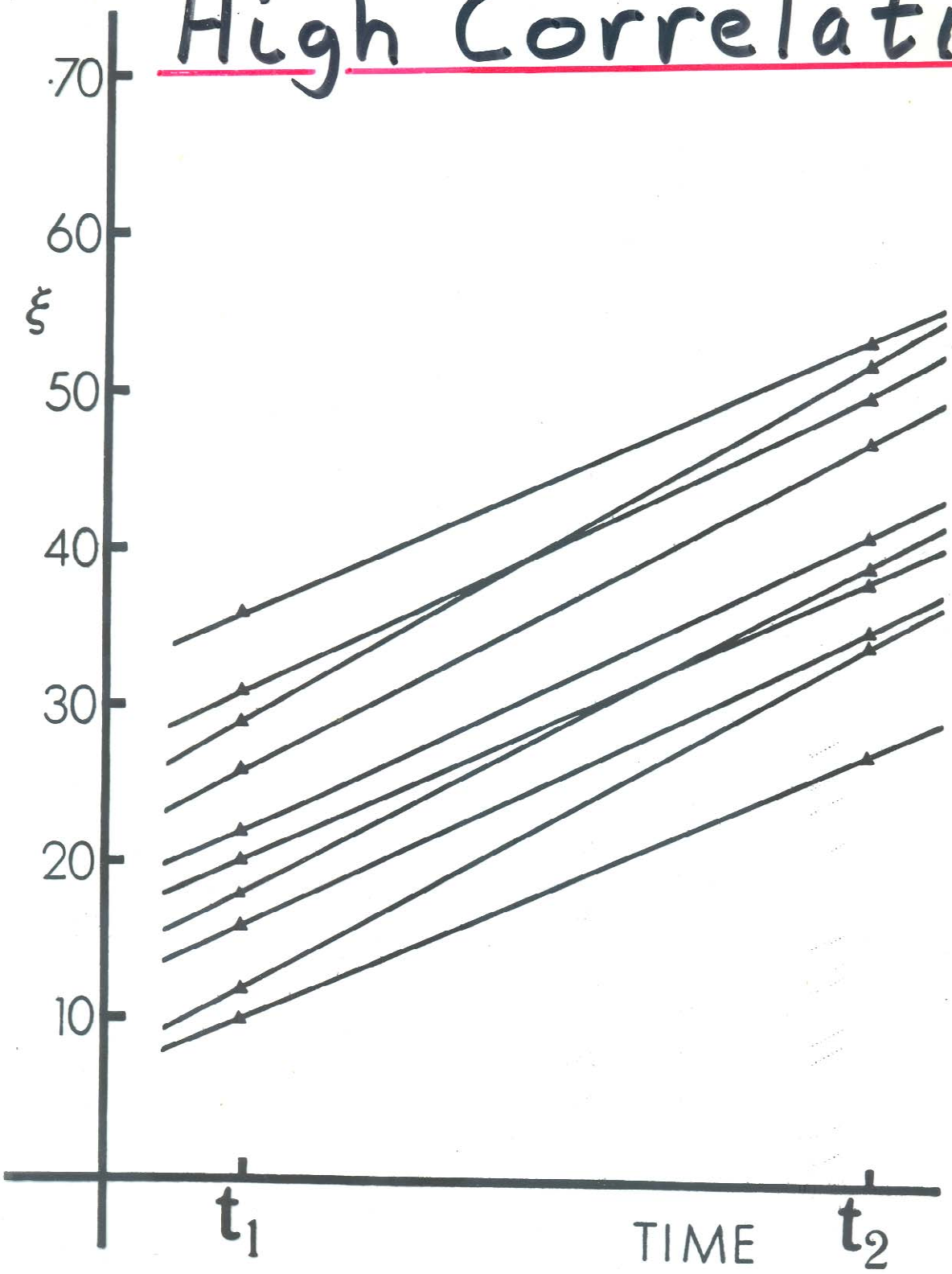
MESSAGE

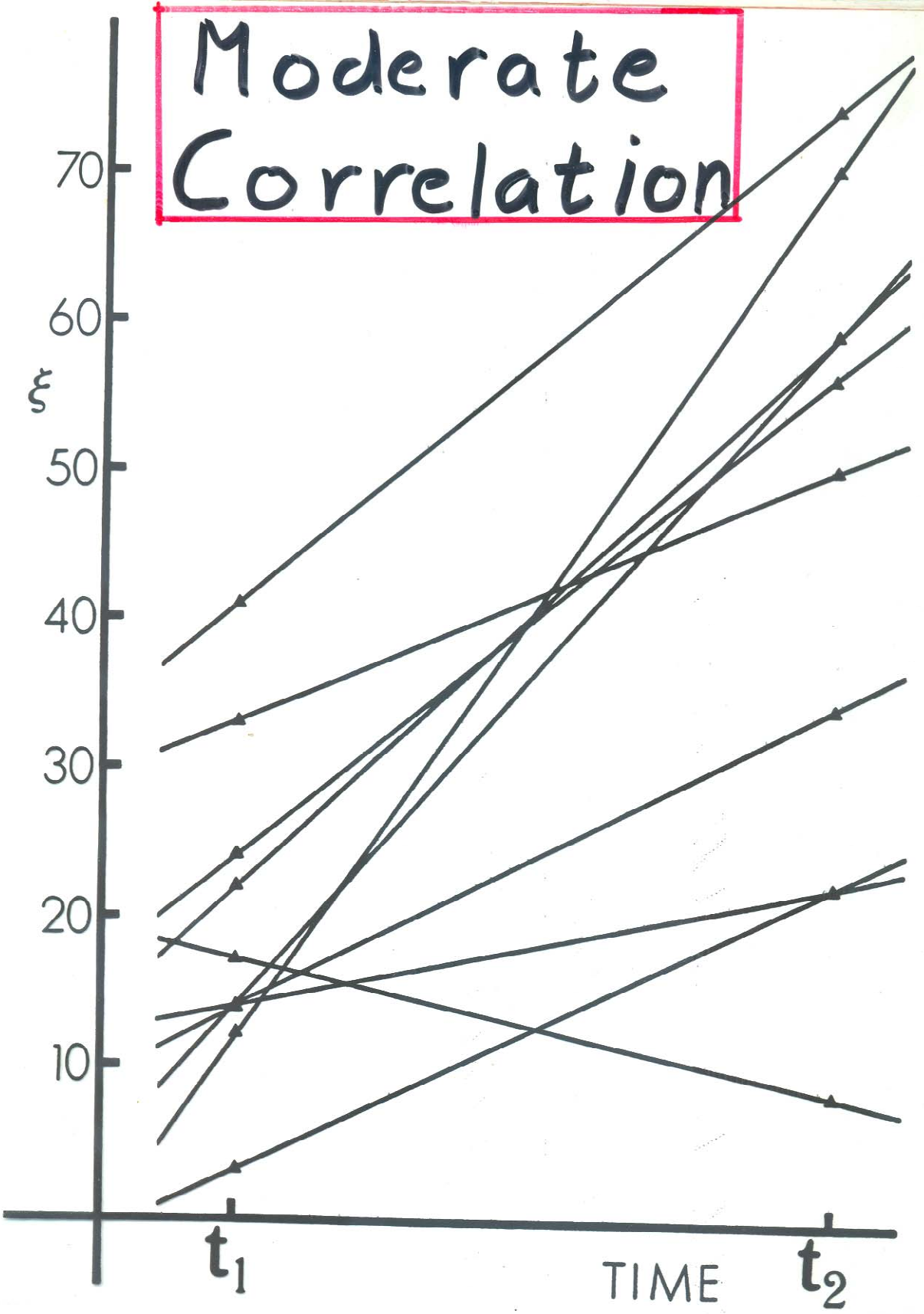
THE DIFFERENCE SCORE IS RELIABLE WHEN INDIVIDUAL DIFFERENCES IN TRUE CHANGE EXIST.

$$\rho(D) = \frac{\sigma_{\Delta}^2}{\sigma_{\Delta}^2 + \sigma_{\epsilon_2 - \epsilon_1}^2}$$

Reliability is not accuracy or precision: see shoe-shopping example

High Correlation




Moderate Correlation
70
60
ξ
50
40
30
20
10
t1
TIME
t2

VALUES OF $\rho(D)/\bar{\rho}(x)$ AND $\rho(D)$ WHEN $\rho_{\xi_1 D} = 0$ AND $\rho(x_2) = .9$

	$\rho(x_1)$		
$\rho_{\xi_1 \xi_2}$	.6	.7	.8
	$\rho(D)/\bar{\rho}(x)$		
.4	1.06	1.03	1.00
.6	.86	.88	.90
.8	.53	.60	.67
	$\rho(D)$		
.4	.79	.82	.85
.6	.65	.71	.76
.8	.40	.48	.57

VALUES OF $\rho(D)/\rho(X)$ AND $\rho(D)$ WHEN $\sigma_{X_1} = \sigma_{X_2}$ AND $\rho(X_1) = \rho(X_2)$

		$\rho(X)$		
$\rho_{\xi_1\xi_2}$	$\rho_{\xi_1\Delta}$	.7	.8	.9
		$\rho(D)/\rho(X)$		
.4	−.55	.83	.88	.94
.6	−.45	.69	.77	.87
.8	−.32	.45	.55	.71
		$\rho(D)$		
.4	−.55	.58	.71	.84
.6	−.45	.48	.62	.78
.8	−.32	.32	.44	.64

$$D_p = X_{2p} - X_{1p}$$

IS AN UNBIASED ESTIMATE OF

$$\Delta_p = \xi_{2p} - \xi_{1p} .$$

HOW CAN AN UNBIASED ESTIMATE BE UNFAIR?

You can determine
from
(the correlation matrix for)
the longitudinal data
whether (or not)
you are measuring
the same thing
over time

Myths Companion

STAT 222
week 2
D Rogosa

M1 individual growth $\xi_p(t) = f(\xi, t)$
proportional growth to asymptote λ_p
Amount of Change $\xi_p(t) = \lambda_p - (\lambda_p - \xi_p(0))\, e^{-\gamma_p t}$
$A_p(t, t+c) = [\lambda_p - \xi_p(0)][1 - e^{-\gamma_p c}]\, e^{-\gamma_p(t)}$ trading places

M2 $X_{ip} = \xi_{ip} + \varepsilon$ (fallible score, see back)
$D = X_2 - X_1$, $E(D) = \xi_2 - \xi_1$
$\sigma^2_\Delta = (t_2 - t_1)^2 \sigma^2_\omega$ individual differences in change
reliability $\rho(D) = \dfrac{\sigma^2_\Delta}{\sigma^2_\Delta + \sigma^2_{\varepsilon_2} - \varepsilon_1}$

M3 $\rho_{\xi(t)\,\xi(t+c)}$ on back, depends on choice of t, $t+c$
for $\rho_{\xi, \Delta} = 0$ $\rho_{\xi_1 \xi_2} = (1 + \sigma^2_\Delta / \sigma^2_{\xi_1})^{-1/2}$

M4 $\rho_{\xi(t_1)\omega}$ on back depends on $t_1 - t^0$
Bias of $r_{x,D}$ $E(r_{x,D}) = \rho_{\xi(t_1)\omega} \sqrt{\rho(x_1)\rho(D)} - \dfrac{\sigma^2_{\varepsilon_1} - \sigma_{\varepsilon_1 \varepsilon_2}}{\sigma_{x_1} \sigma_D}$
proportional bias; additional downward

M5 Standardized Tautology
$\dfrac{E(\xi_2 \mid \xi_1 = c) - \mu_{\xi_2}}{\sigma_{\xi_2}} < \dfrac{c - \mu_{\xi_1}}{\sigma_{\xi_1}} \Rightarrow \rho_{\xi_1 \xi_2} < 1$

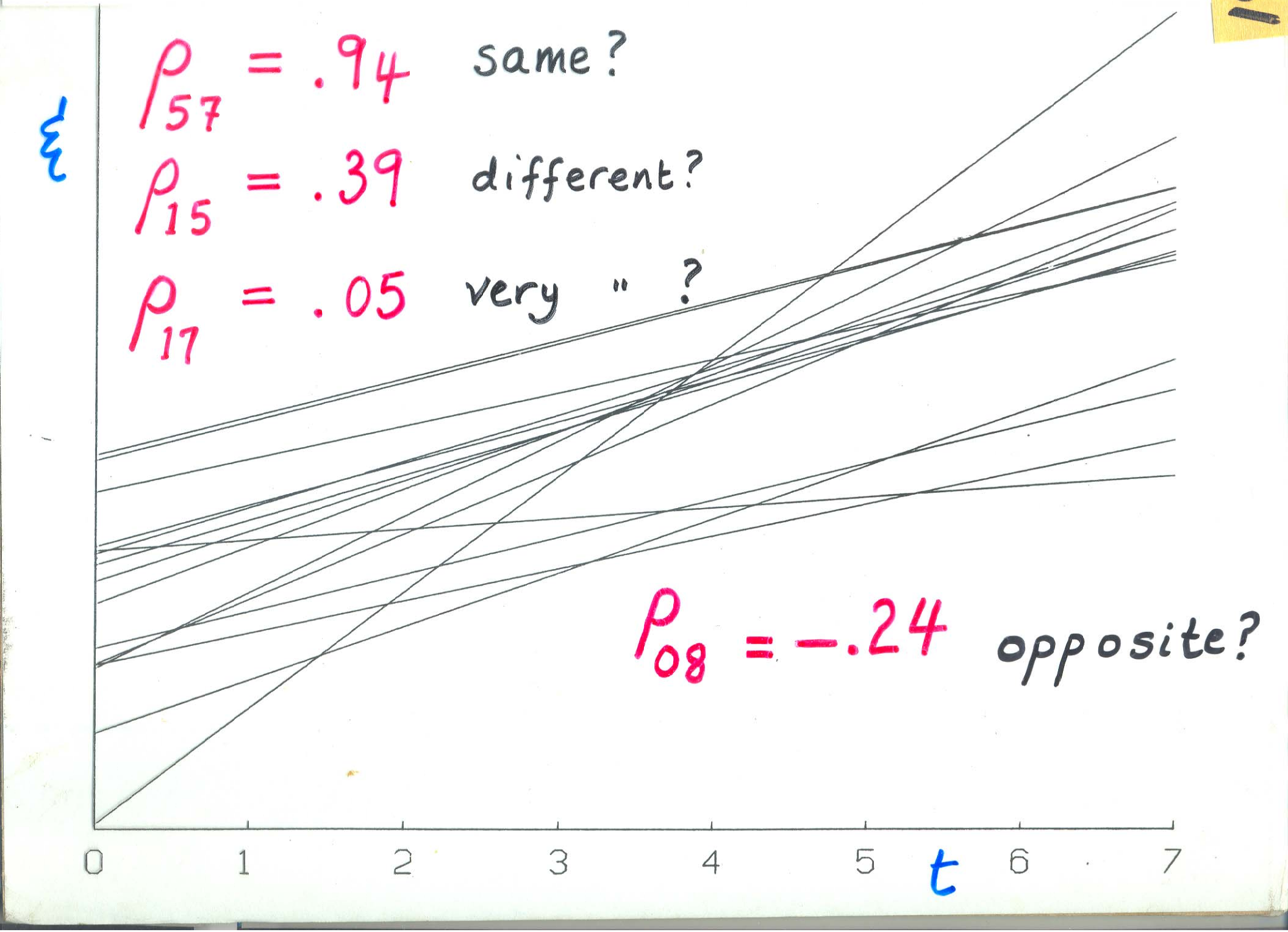
In metric of data $E(\xi_2 \mid \xi_1 = c) - \mu_{\xi_2} < c - \mu_{\xi_1}$
$\Rightarrow \rho_{\xi_1, \Delta} < 0$

M6/7 resid change in sample $X_2 \cdot X_1$ bias poor reliability
Correlation $\rho[\Delta \cdot \xi(t_1)]\,\omega$ vs $\rho\omega$
$\Delta \cdot \xi(t_1) = \xi(t_2) \cdot \xi(t_1)$

LARGE INDIVIDUAL DIFFERENCES IN GROWTH LOWER THE BETWEEN-WAVE CORRELATIONS

EXAMPLE: FOR $\rho_{\xi_1 \Delta} = 0$

$$\rho_{\xi_1 \xi_2} = \frac{1}{\sqrt{1 + \sigma_\Delta^2 / \sigma_{\xi_1}^2}}$$


$\rho_{57} = .94$ same?
$\rho_{15} = .39$ different?
$\rho_{17} = .05$ very " ?
$\rho_{08} = -.24$ opposite?
ξ
0
1
2
3
4
5
t
6
7

The correlation between change and initial status is

a) Negative

b) Zero

c) Positive

---

d) All of the above

Myths Companion

STAT 222
week 2
D Rogosa

M1 individual growth $\xi_p(t) = f(\xi, t)$
proportional growth to asymptote λ_p
Amount of Change $\xi_p(t) = \lambda_p - (\lambda_p - \xi_p(0)) e^{-\gamma_p t}$
$A_p(t, t+c) = [\lambda_p - \xi_p(0)][1 - e^{-\gamma_p c}] e^{-\gamma_p(t)}$ trading places

M2 $X_{ip} = \xi_{ip} + \varepsilon$ (fallible score, see back)
$D = X_2 - X_1$, $E(D) = \xi_2 - \xi_1$
$\sigma_\Delta^2 = (t_2 - t_1)^2 \sigma_\theta^2$ individual differences in change
reliability $\rho(D) = \dfrac{\sigma_\Delta^2}{\sigma_\Delta^2 + \sigma_{\varepsilon_2 - \varepsilon_1}^2}$

M3 $\rho_{\xi(t)\,\xi(t+c)}$ on back, depends on choice of t, $t+c$
for $\rho_{\xi_1 \Delta} = 0$ $\rho_{\xi_1 \xi_2} = (1 + \sigma_\Delta^2 / \sigma_{\xi_1}^2)^{-1/2}$

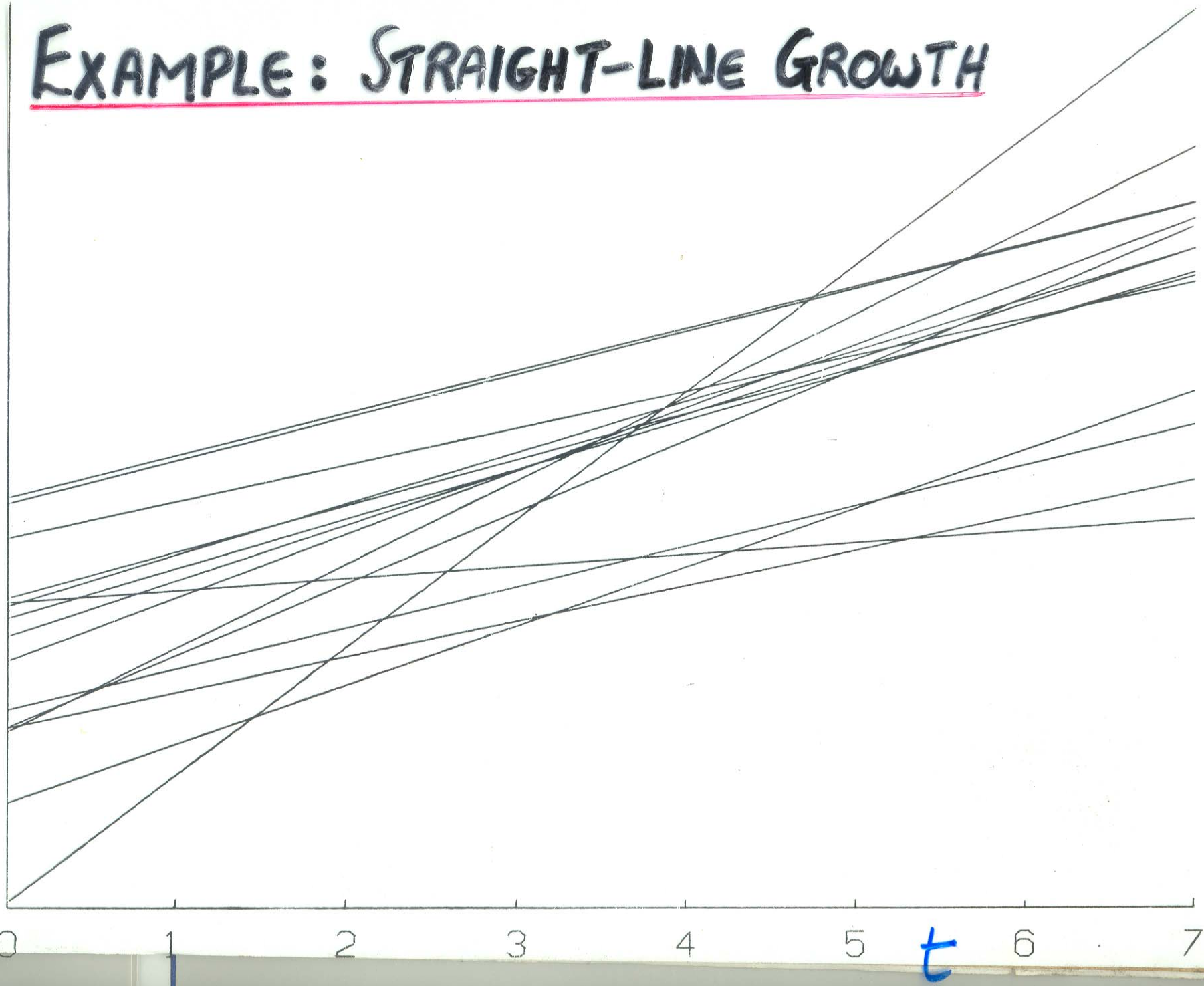
M4 $\rho_{\xi(t_1)\theta}$ on back depends on $t_1 - t^0$
Bias of $r_{X_1 D}$ $E(r_{X_1 D}) = \rho_{\xi(t_1)\theta} \sqrt{\rho(X_1)\rho(D)} - \dfrac{\sigma_{\varepsilon_1}^2 - \sigma_{\varepsilon_1 \varepsilon_2}}{\sigma_{X_1} \sigma_D}$
proportional bias; additional downward

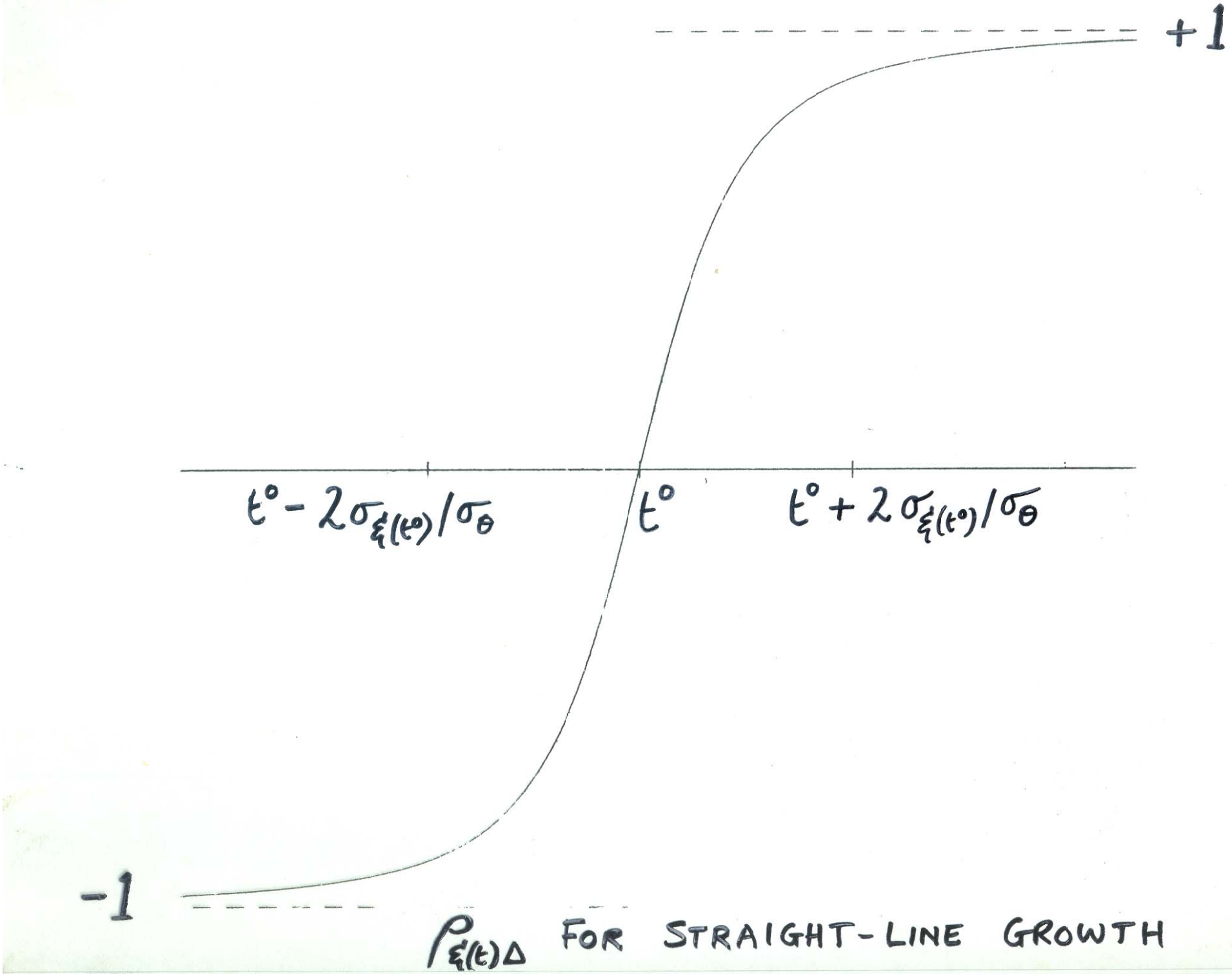
M5 Standardized Tautology
$\dfrac{E(\xi_2 \mid \xi_1 = c) - \mu_{\xi_2}}{\sigma_{\xi_2}} < \dfrac{c - \mu_{\xi_1}}{\sigma_{\xi_1}} \Rightarrow \rho_{\xi_1 \xi_2} < 1$

In metric of data $E(\xi_2 \mid \xi_1 = c) - \mu_{\xi_2} < c - \mu_{\xi_1}$
$\Rightarrow \rho_{\xi_1 \Delta} < 0$

M6/7 resid change in sample $X_2 \cdot X_1$ bias poor reliability
Correlation $\rho[\Delta \cdot \xi(t_1)]\,\omega$ vs $\rho\,\omega$
$\Delta \cdot \xi(t_1) = \xi(t_2) \cdot \xi(t_1)$

EXAMPLE: STRAIGHT-LINE GROWTH





$\rho_{\xi(t)\Delta}$ FOR STRAIGHT-LINE GROWTH

Correlation between Change and Initial Status

t_I	$\rho_{\xi(t_I)\Delta}$	$\rho_{X_i(X_{i+c}-X_i)}$ c=1	c=3
0	-.71	-.50	-.69
1	-.55	-.48	-.59
2	-.32	-.44	-.47
3	0	-.36	-.29
4	.32	-.25	.00
5	.55	-.12	.17
6	.71	-.02	.30
7	.80	.02	.42

YOU CAN'T AVOID REGRESSION TOWARD THE MEAN

Myths Companion

STAT 222
week 2
D Rogosa

M1 individual growth $\xi_p(t) = f(\xi, t)$
proportional growth to asymptote λ_p
Amount of Change $\xi_p(t) = \lambda_p - (\lambda_p - \xi_p(0))\, e^{-\gamma_p t}$
$A_p(t, t+c) = [\lambda_p - \xi_p(0)][1 - e^{-\gamma_p c}]\, e^{-\gamma_p(t)}$ trading places

---

M2 $X_{ip} = \xi_{ip} + \varepsilon$ (fallible score, see back)
$D = X_2 - X_1$, $E(D) = \xi_2 - \xi_1$
$\sigma_\Delta^2 = (t_2 - t_1)^2 \sigma_\Theta^2$ individual differences in change
reliability $\rho(D) = \dfrac{\sigma_\Delta^2}{\sigma_\Delta^2 + \sigma_{\varepsilon_2 - \varepsilon_1}^2}$

M3 $\rho_{\xi(t)\,\xi(t+c)}$ on back, depends on choice of t, $t+c$
for $\rho_{\xi_1 \Delta} = 0$ $\rho_{\xi_1 \xi_2} = \left(1 + \sigma_\Delta^2 / \sigma_{\xi_1}^2\right)^{-1/2}$

M4 $\rho_{\xi(t_1)\Theta}$ on back depends on $t_1 - t^0$
Bias of $r_{X_1 D}$ $E(r_{X_1 D}) = \rho_{\xi(t_1)\Theta} \sqrt{\rho(X_1)\rho(D)} - \dfrac{\sigma_{\varepsilon_1}^2 - \sigma_{\varepsilon_1 \varepsilon_2}}{\sigma_{X_1}\sigma_D}$
proportional bias; additional downward

M5 Standardized Tautology
$\dfrac{E(\xi_2 \mid \xi_1 = c) - \mu_{\xi_2}}{\sigma_{\xi_2}} < \dfrac{c - \mu_{\xi_1}}{\sigma_{\xi_1}} \Rightarrow \rho_{\xi_1 \xi_2} < 1$

In metric of data $E(\xi_2 \mid \xi_1 = c) - \mu_{\xi_2} < c - \mu_{\xi_1}$
$\Rightarrow \rho_{\xi_1 \Delta} < 0$

M6/7 resid change in sample $X_2 - X_1$ bias poor reliability
Correlation $\rho[\Delta \cdot \xi(t_1)]\, \omega$ vs $\rho\, \omega$
$\Delta \cdot \xi(t_1) = \xi(t_2) \cdot \xi(t_1)$

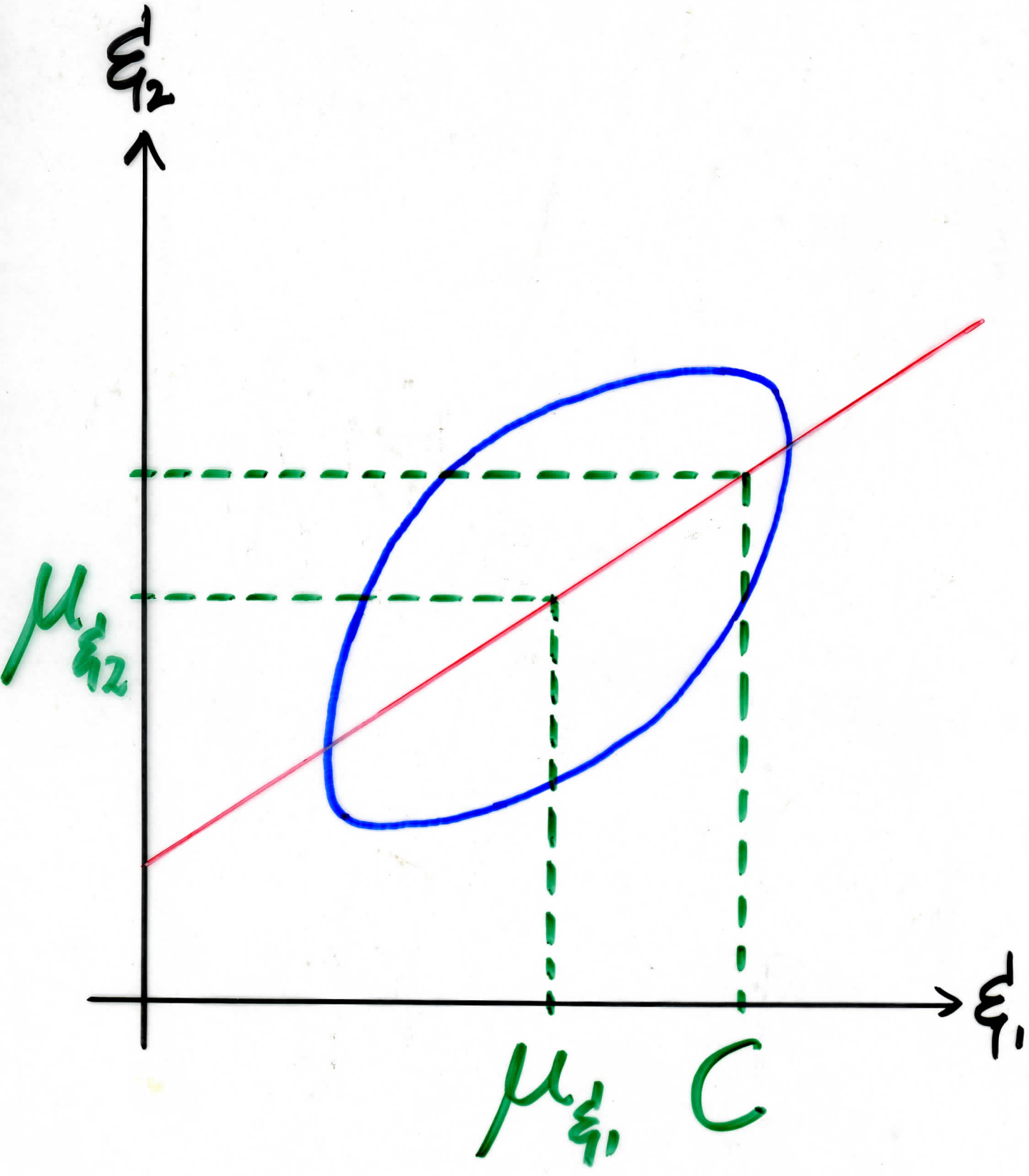
Regression Toward The Mean

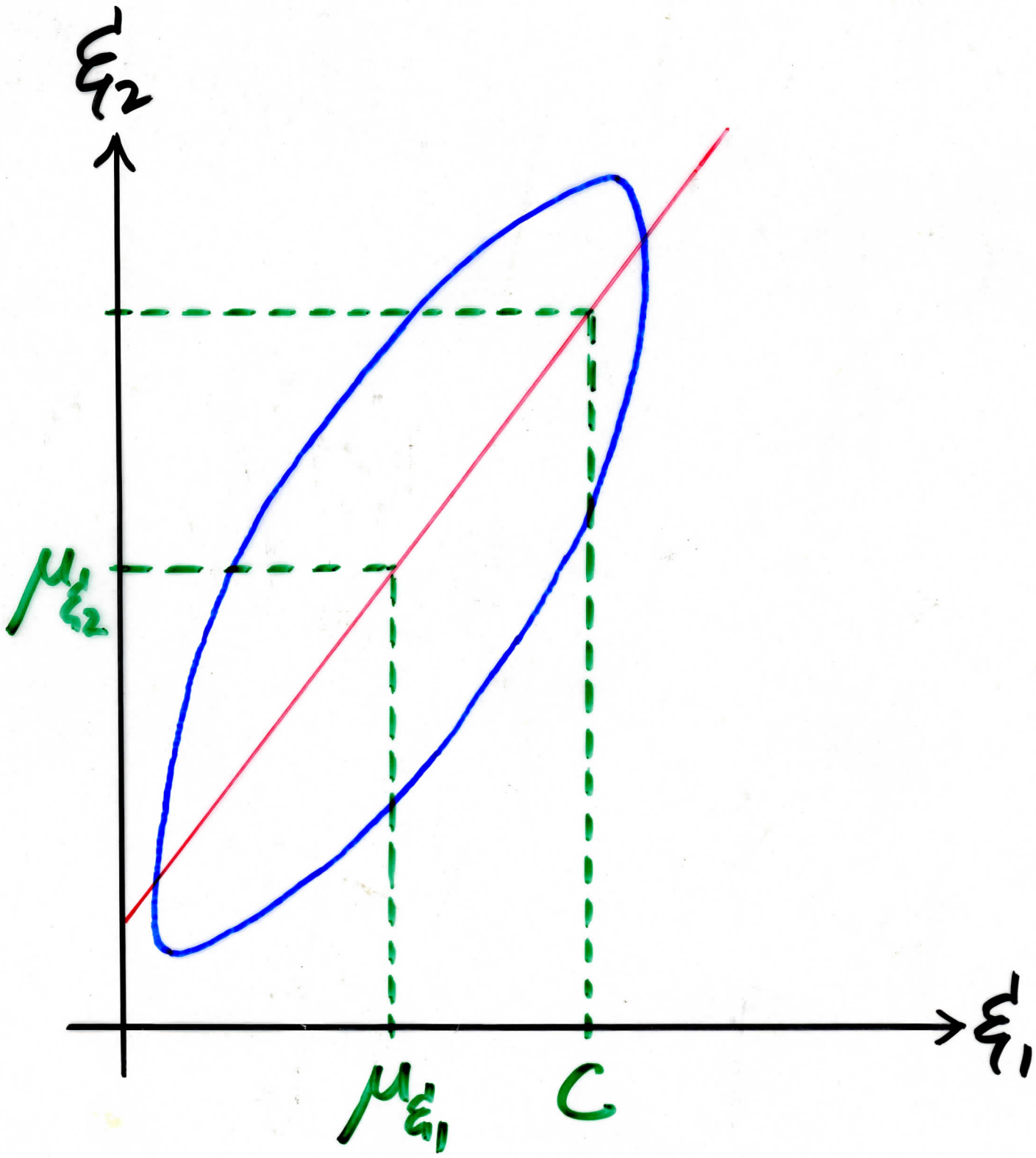
$$\frac{E(\xi_2 \mid \xi_1 = c) - \mu_{\xi_2}}{\sigma_{\xi_2}} < \frac{c - \mu_{\xi_1}}{\sigma_{\xi_1}}$$

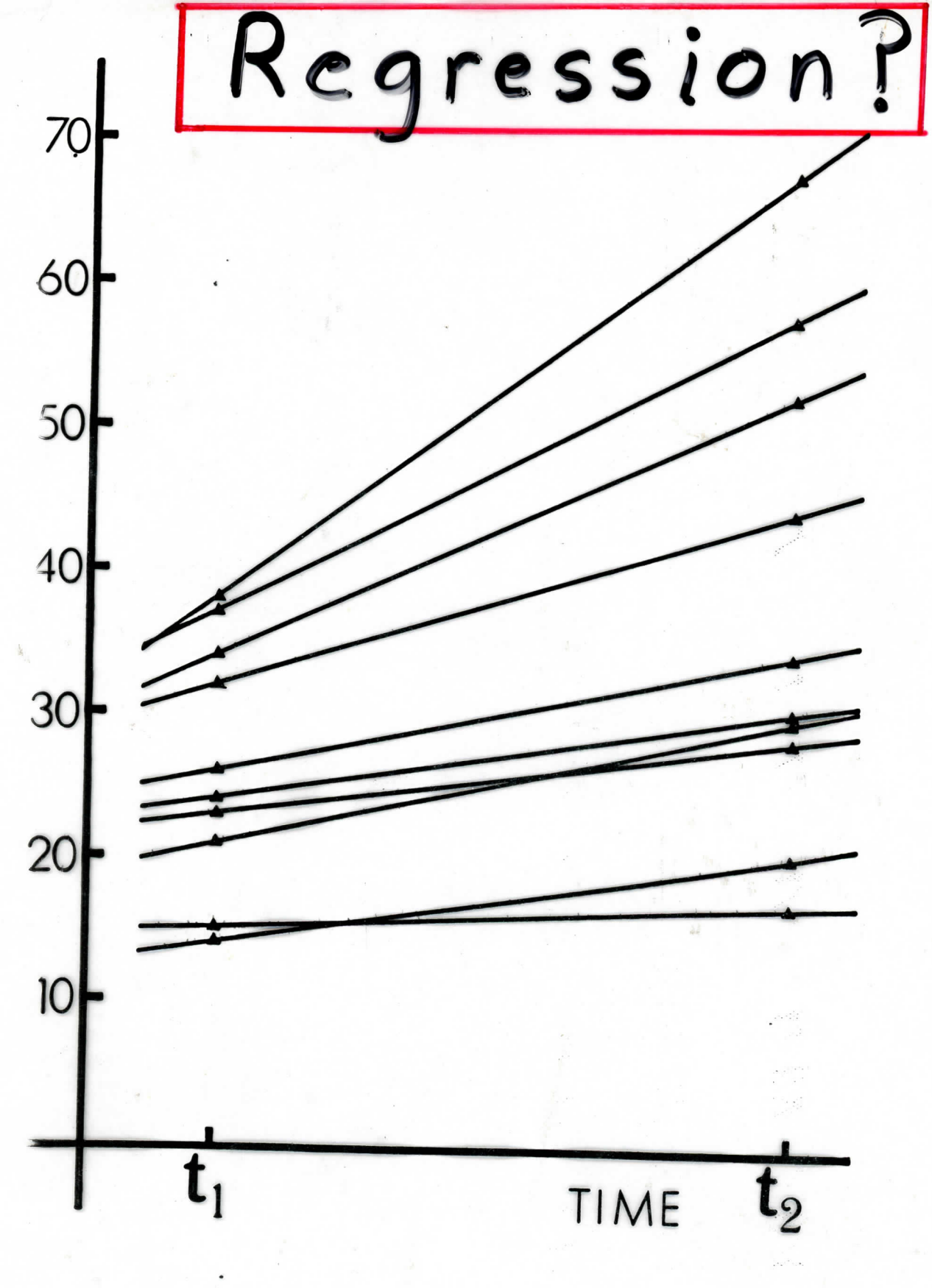
$$\Rightarrow \rho_{\xi_1 \xi_2} < 1$$

$$E(\xi_2 \mid \xi_1 = c) - \mu_{\xi_2} < c - \mu_{\xi_1}$$

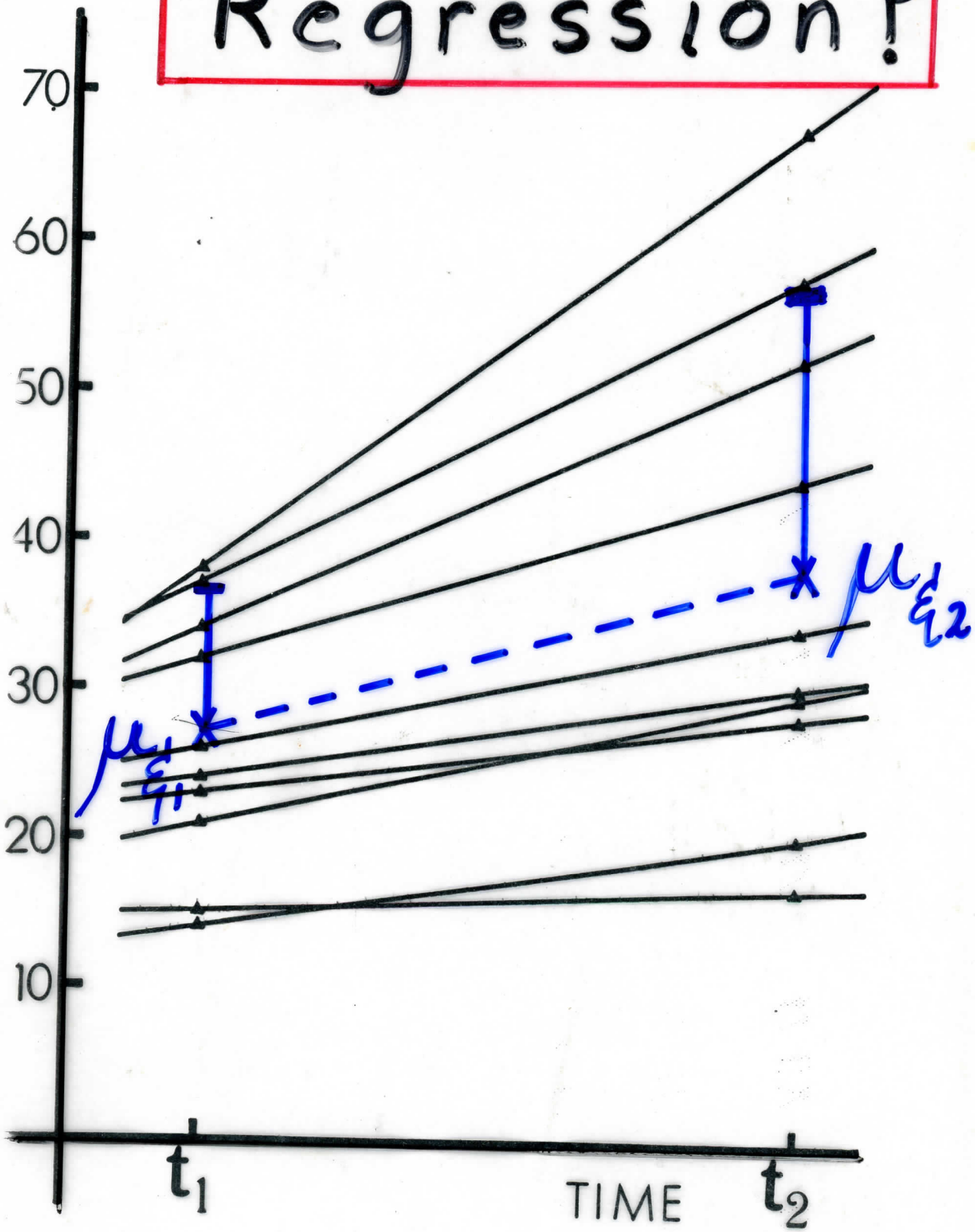
$$\Rightarrow \rho_{\xi_1 \Delta} < 0$$


ξ_2
μ_{ξ_2}
ξ_1
μ_{ξ_1}
C


ξ_2
μ_{ξ_2}
μ_{ξ_1}
C
ξ_1


Regression?
70
60
50
40
30
20
10
t_1
TIME
t_2

Regression?



Rx residual
change measures
cure
what ails
the difference
score

Residual Change As A Measure Of Change

True Residual Change

$$\xi_{2p} - \mu_{\xi_2} - \beta_{\xi_2\xi_1}(\xi_{1p} - \mu_{\xi_1})$$

Usual Sample Estimate

$$\hat{R}_p = X_{2p} - \bar{X}_2 - \hat{\beta}_{X_2X_1}(X_{1p} - \bar{X}_1)$$

Properties of $\hat{R}$

Bias?	Yes
Precision?	Poor
Reliability?	$\approx \rho(D)$
Logic?	How much would Person p have changed if everyone had started out equal?

Myths Companion

STAT 222
week 2
D Rogosa

M1 individual growth $\xi_p(t) = f(\xi, t)$
proportional growth to asymptote λ_p
Amount of Change $\xi_p(t) = \lambda_p - (\lambda_p - \xi_p(0))\, e^{-\gamma_p t}$
$A_p(t, t+c) = [\lambda_p - \xi_p(0)][1 - e^{-\gamma_p c}]\, e^{-\gamma_p(t)}$ trading places

---

M2 $X_{ip} = \xi_{ip} + \varepsilon$ (fallible score, see back)
$D = X_2 - X_1$, $E(D) = \xi_2 - \xi_1$

$\sigma^2_\Delta = (t_2 - t_1)^2 \sigma^2_\omega$ individual differences in change

reliability $\rho(D) = \dfrac{\sigma^2_\Delta}{\sigma^2_\Delta + \sigma^2_{\varepsilon_2} - \varepsilon_1}$

M3 $\rho_{\xi(t)\,\xi(t+c)}$ on back, depends on choice of t, $t+c$

for $\rho_{\xi,\Delta} = 0$ $\quad \rho_{\xi_1 \xi_2} = (1 + \sigma^2_\Delta / \sigma^2_{\xi_1})^{-1/2}$

M4 $\rho_{\xi(t_1)\omega}$ on back depends on $t_1 - t^0$

Bias of $r_{X,D}$ $\quad E(r_{X,D}) = \rho_{\xi(t_1)\omega} \sqrt{\rho(X_1)\rho(D)} - \dfrac{\sigma^2_{\varepsilon_1} - \sigma_{\varepsilon_1 \varepsilon_2}}{\sigma_{X_1}\sigma_D}$

proportional bias; additional downward

M5 Standardized Tautology

$$\frac{E(\xi_2 \mid \xi_1 = c) - \mu_{\xi_2}}{\sigma_{\xi_2}} < \frac{c - \mu_{\xi_1}}{\sigma_{\xi_1}} \Rightarrow \rho_{\xi_1 \xi_2} < 1$$

In metric of data $E(\xi_2 \mid \xi_1 = c) - \mu_{\xi_2} < c - \mu_{\xi_1}$

$\Rightarrow \rho_{\xi,\Delta} < 0$

M6/7 resid change in sample $X_2 - X_1$ biased poor reliability

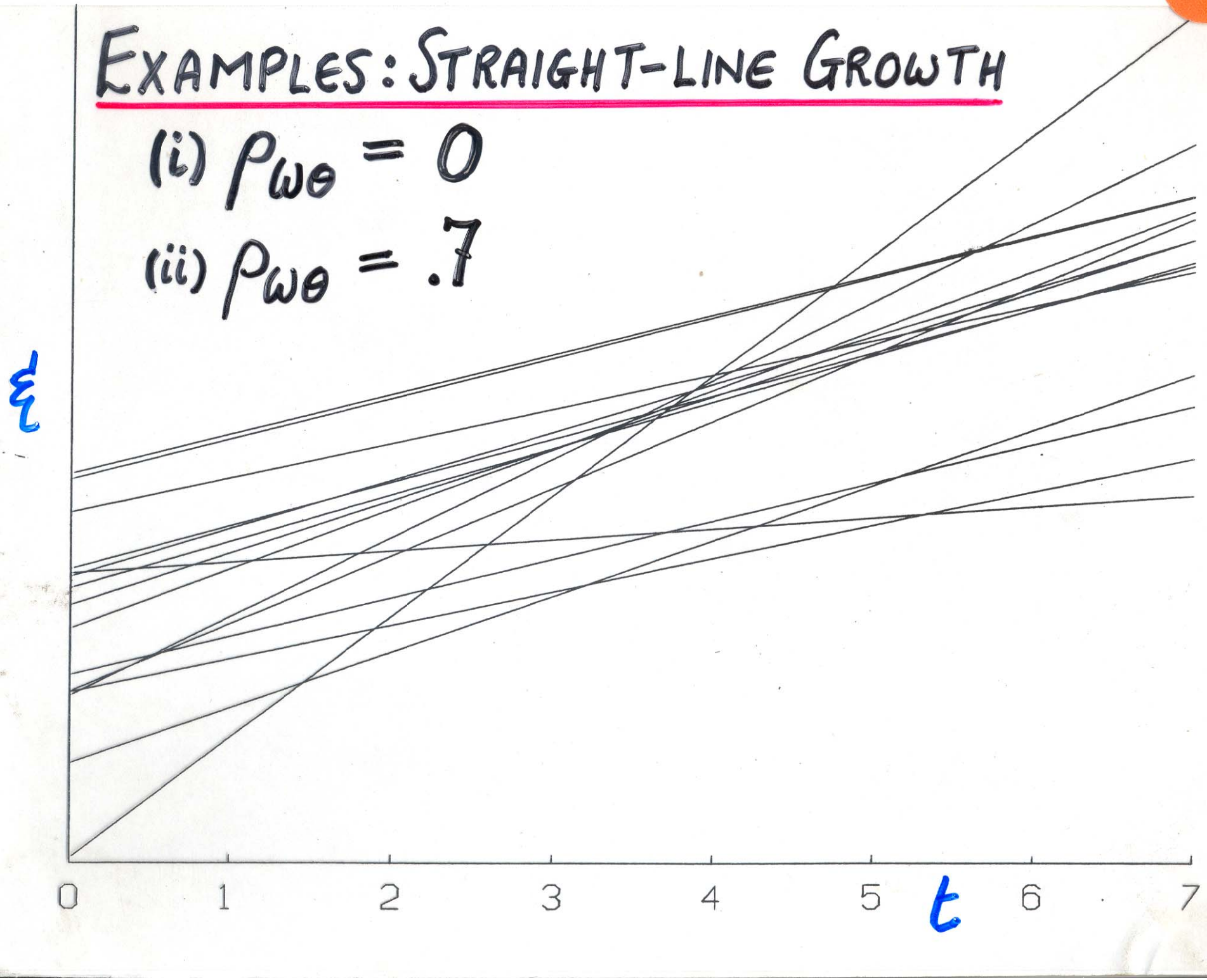
Correlation $\rho[\Delta \cdot \xi(t_1)]\omega$ vs $\rho\omega$

$\Delta \cdot \xi(t_1) = \xi(t_2) \cdot \xi(t_1)$

EXAMPLES: STRAIGHT-LINE GROWTH

(i) $\rho_{\omega\theta} = 0$

(ii) $\rho_{\omega\theta} = .7$

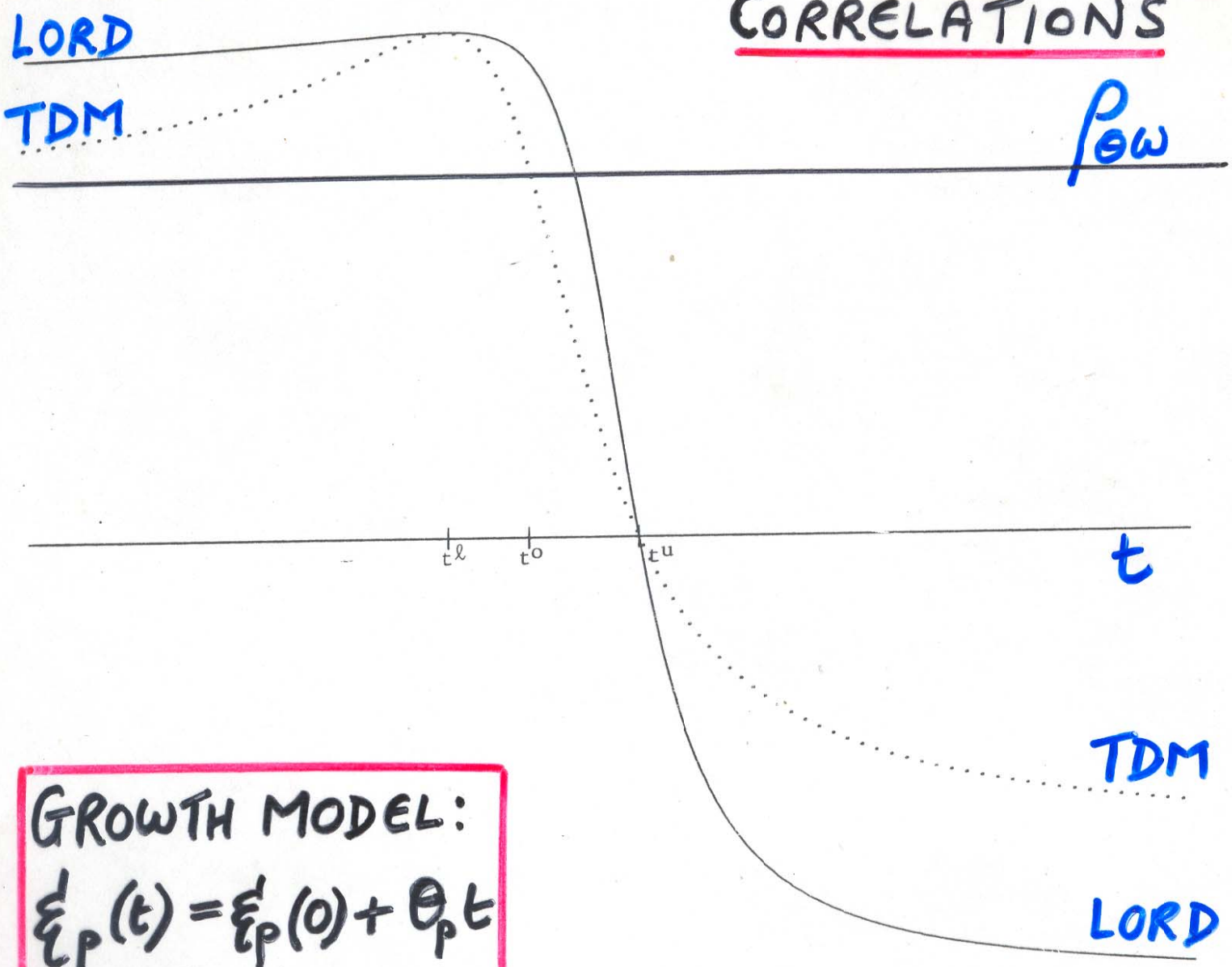


Residual Change And Correlates of Change

Parameter	Sample Quantity
TDM (1966)	
$\rho_{[\xi(t_2)\cdot\xi(t_1)]\omega} = \rho_{[\Delta\cdot\xi(t_1)]\omega}$	$r_{\hat{R}W}$
LORD (1958)	
$\rho_{\Delta\omega\cdot\xi(t_1)} = \rho_{\xi(t_2)\omega\cdot\xi(t_1)} = \rho_{[\xi(t_2)\cdot\xi(t_1)]\omega\cdot\xi(t_1)}$	$r_{X_2W\cdot X_1}$

Do These Quantities Reflect Systematic Individual Differences in Growth?

FUNCTIONAL FORM OF RESIDUAL CHANGE CORRELATIONS



GROWTH MODEL:

$$\xi'_p(t) = \xi'_p(0) + \theta_p t$$

Values of $\rho_{[\xi'(t_2)\cdot\xi'(t_1)]W}$ (TDM)

t_1	$\rho_{W\theta}=0$	$\rho_{W\theta}=.7$
0	.64	.92
1	.50	.92
2	.29	.85
3	0	.70
4	−.29	.47
5	−.50	.25
6	−.64	.07
7	−.73	−.06
8	−.78	−.15

THE WAY TO INVESTIGATE CORRELATES OF CHANGE IS TO MODEL INDIVIDUAL GROWTH AND THEN INVESTIGATE SYSTEMATIC INDIVIDUAL DIFFERENCES IN GROWTH

MORAL

CANNOT ASSESS CORRELATES OF CHANGE BY IGNORING INDIVIDUAL GROWTH.

Longitudinal Reasons to AVOID Structural Equation Models

David Rogosa

Stanford University

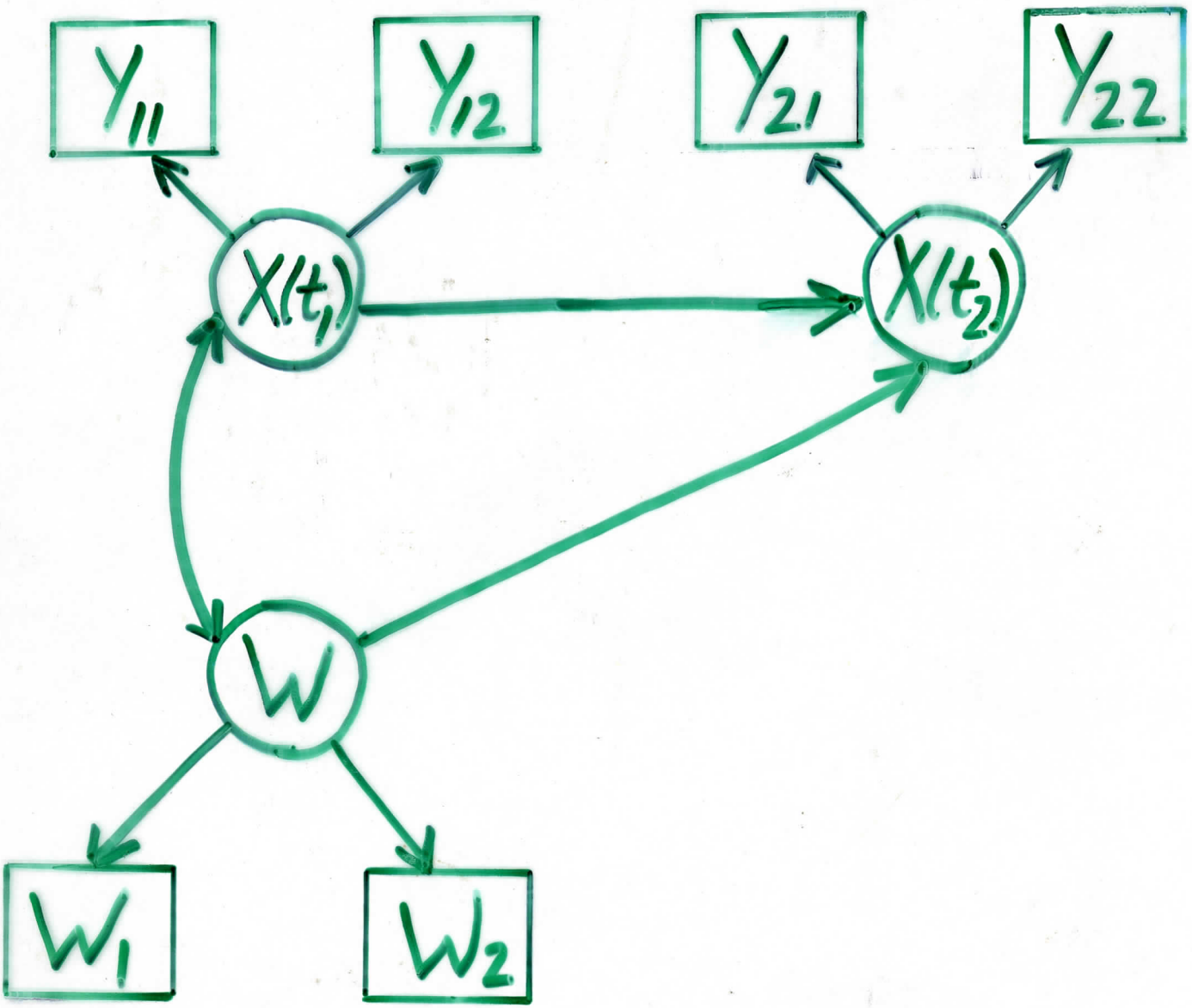
rag@leland.stanford.edu

School of Education
Stanford CA 94305

Quantitative Social Science Seminar Series
U.C. Berkeley, March 14, 1994

2. Structural Equation Models

Example: Stability / Change in Alienation (Jöreskog, 19**)



Structural equation

$$X_2 = \alpha_0 + \beta_1 X_1 + \beta_2 W$$

Exogenous influence on change

$$\beta_2 = \beta_{X(t_2)\, W \cdot X(t_1)}$$

Jöreskog:

$\hat{\beta}_2 < 0 \Rightarrow$ "high SES reduces alienation"

Properties (Moments of Observables) of Collections of Growth Curves

STAT 222 Week 2 D Rogosa

for indv p $\quad \xi_p(t) = \xi_p(0) + \theta_p t$ $\quad t_i \ (i = 1, \ldots T)$ $\quad p \ (p = 1, \ldots n)$

"centering"

$$\xi_p(t) = \xi_p(t^0) + \theta_p (t - t^0)$$

centering, scale

$t^0 = -\sigma_{\xi(0)\theta} / \sigma^2_\theta$

$\rho_{\xi(t^0)\theta} = 0$, min Var($\xi$)

Scale $K = \sigma_{\xi(t^0)} / \sigma_\theta$

time metric

Moments

Covariance $\sigma_{\xi(t_1)\xi(t_2)} =$

$$(t_1 - t^0)(t_2 - t^0)\sigma^2_\theta + \sigma^2_{\xi(t^0)}$$

Variance $\sigma^2_{\xi(t)} = \sigma^2_{\xi(t^0)} + ((t - t^0)/K)^2 \sigma^2_{\xi(t^0)}$

$$\sigma^2_{\xi(t)} / \sigma^2_{\xi(t^0)} = 1 + \left(\frac{t - t^0}{K}\right)^2$$

Correl Change, initial status $\rho_{\theta\xi(t)} = \dfrac{t - t^0}{[K^2 + (t - t^0)^2]^{1/2}}$

exogenous var w

$$\rho_{w\xi(t)} = \frac{(t - t^0)\rho_{w\theta} + K\rho_{w\xi(t^0)}}{[K^2 + (t - t^0)^2]^{1/2}}$$

where $t^u = t^0 + K\left(\dfrac{\rho_{w\theta}}{\rho_{w\xi(t^0)}}\right)$ $\quad t^l = t^0 - K\left(\dfrac{\rho_{w\xi(t^0)}}{\rho_{w\theta}}\right)$

Week 1 example: (p. 64) $\theta \sim U[1, 9]$, $\xi(t^0) \sim U[38, 62]$

$t^0 = 2$ $\quad \sigma^2_\theta = 5.333$ $\quad \sigma^2_{\xi(t^0)} = 48$ $\quad \rho_{w\theta} = 0$ $\quad \rho_{w\xi(t^0)} =$

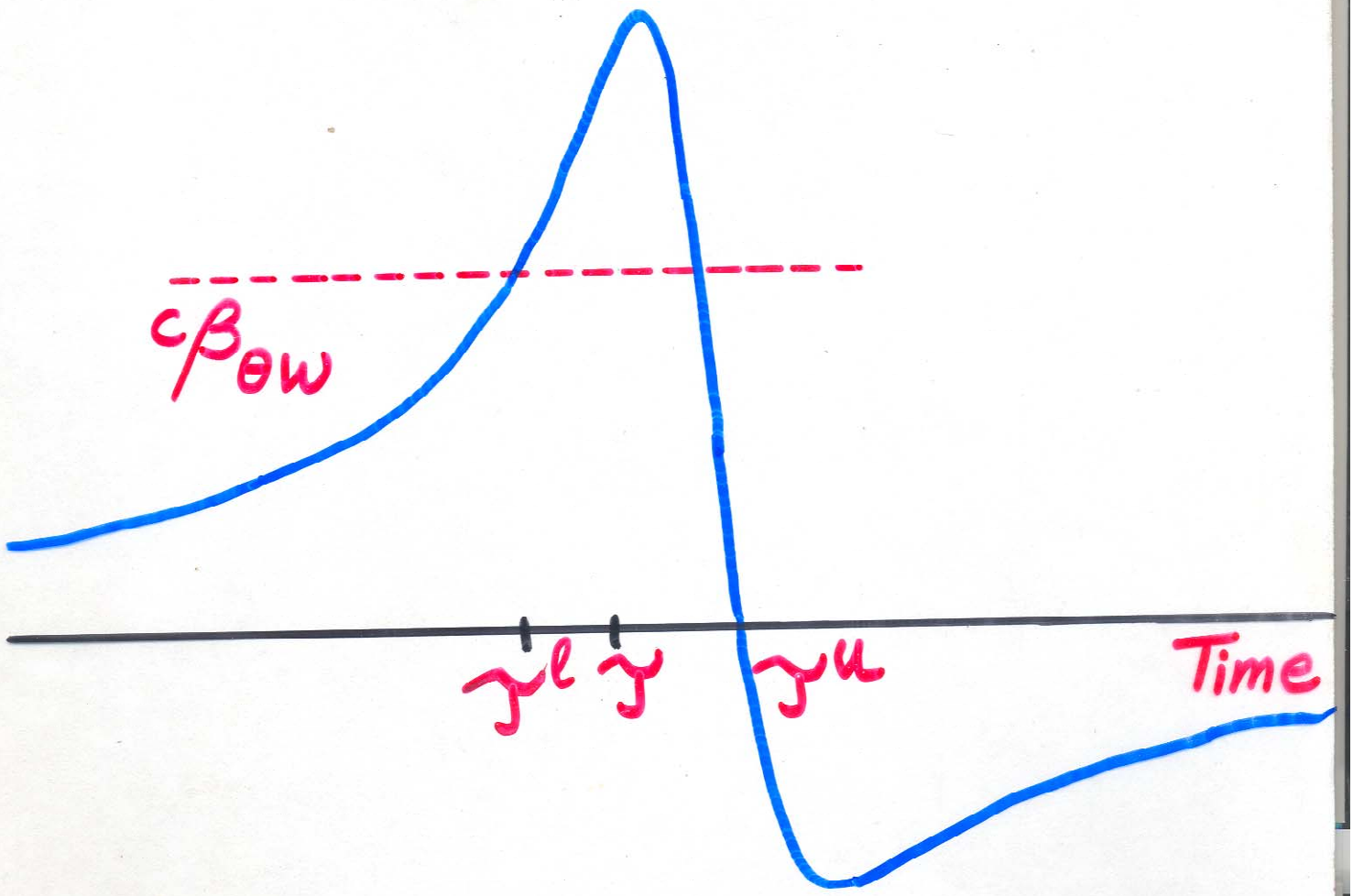
---

at time t_i $\quad X_{ip} = \xi_{ip} + \varepsilon$ $\quad \varepsilon \sim (0, \sigma^2_\varepsilon)$ errors in variables

weeks 1 ex $\sigma^2_\varepsilon = 10$

$\beta_{X(t_2)W \cdot X(t_1)}$ for Straight-line Growth

$$\beta_{X(t_2)W \cdot X(t_1)} = \frac{c\left(\frac{\sigma_{X(\tau)}}{\sigma_W}\right)(\tau^u - t)\,\rho_{WX(\tau)}}{(t-\tau)^2 + X^2 - \rho_{W\theta}^2 (t - \tau^l)^2}$$



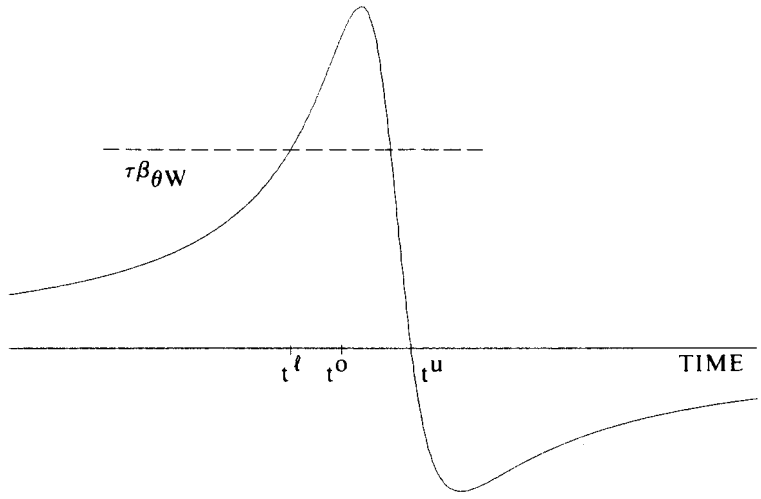


FIGURE 9
Plot of $\beta_{\xi(t+\tau)W\cdot\xi(t)}$ against time for straight-line growth. The value of $\tau\beta_{\theta W}$ is indicated at the dashed line.

respectively. As with the correlational measures, the most salient property of $\beta_{\xi(t_I+\tau)W\cdot\xi(t_I)}$ is the dependence on time of initial status.

For $\rho_{W\theta} = 0$, (25) becomes

$$\beta_{\xi(t+\tau)W\cdot\xi(t)} = \frac{-\tau(t-t^{\circ})\rho_{W\xi(t^{\circ})}[\sigma_{\xi(t^{\circ})}/\sigma_W]}{\kappa^2(1-\rho^2_{W\xi(t^{\circ})}) + (t-t^{\circ})^2}. \tag{27}$$

The function in (27) equals zero for $t = t^{\circ}$ and has minimum and maximum at $t^{\circ} \pm \kappa[1 - \rho^2_{W\xi(t^{\circ})}]^{1/2}$, respectively. Despite the specification that $\beta_{\theta W} = 0$, $\beta_{\xi(t_I+\tau)W\cdot\xi(t_I)}$ will be positive or negative for t_I less than or greater than t°.

Exponential growth. Under the model for exogenous change of Coleman (1968, Equation 11.15), which is rewritten in our (5), the dependence of the regression coefficient on the choice of t_I disappears. Recall that Coleman's model is a special case of systematic individual differences in exponential growth with the restrictions $\gamma_p = \gamma$ and, especially, $\rho_{\lambda W} = 1$. Under these restrictions (which cannot be expected to hold in practice), $\beta_{\xi(t+\tau)W\cdot\xi(t)} = \beta_{\lambda W}(1 - e^{-\gamma\tau})$.

Discussion

At least four purposes for studying change are prominent in the behavioral sciences: (a) the assessment of individual change, (b) the detection of correlates or predictors of change, (c) the comparison of change among experimental groups, and (d) the comparison of change among nonequivalent groups in quasiexperiments (see also Cronbach & Furby, 1970, pp. 77–80). Individual change was the focus of Rogosa et al. (1982), and the present paper moves on to correlates of change. "Understanding Correlates of Change" means how to think about and explicitly formulate systematic individual differences in growth. This understanding (which is achieved by "Modeling Individual Differences in Growth") is a necessary first step in the development of statistical methods to guide the design and analysis of empirical research. A major consequence of this understanding is a call to abandon the teachings of the "Avoid Change at Any Cost" School of Longitudinal Research which have dominated the measurement of change literature. This paper demonstrates that explicit consideration of change—through the parameters of a model for individual growth—is absolutely essential for any serious treatment of correlates of change.

Straight-line Growth Example

$\mathcal{T}=3 \quad \mathcal{K}=3$

Values of $\beta_{X(t_I+5)W \cdot X(t_I)}$

t_I	$\rho_{w\theta}=0$	$\rho_{w\theta}=.7$
0	.85	.70
1	1.05	.85
2	1.15	1.0
3	0	1.2
4	-1.15	1.3
5	-1.05	1.1
6	-.85	.35
7	-.70	-.25
8	-.55	-.5

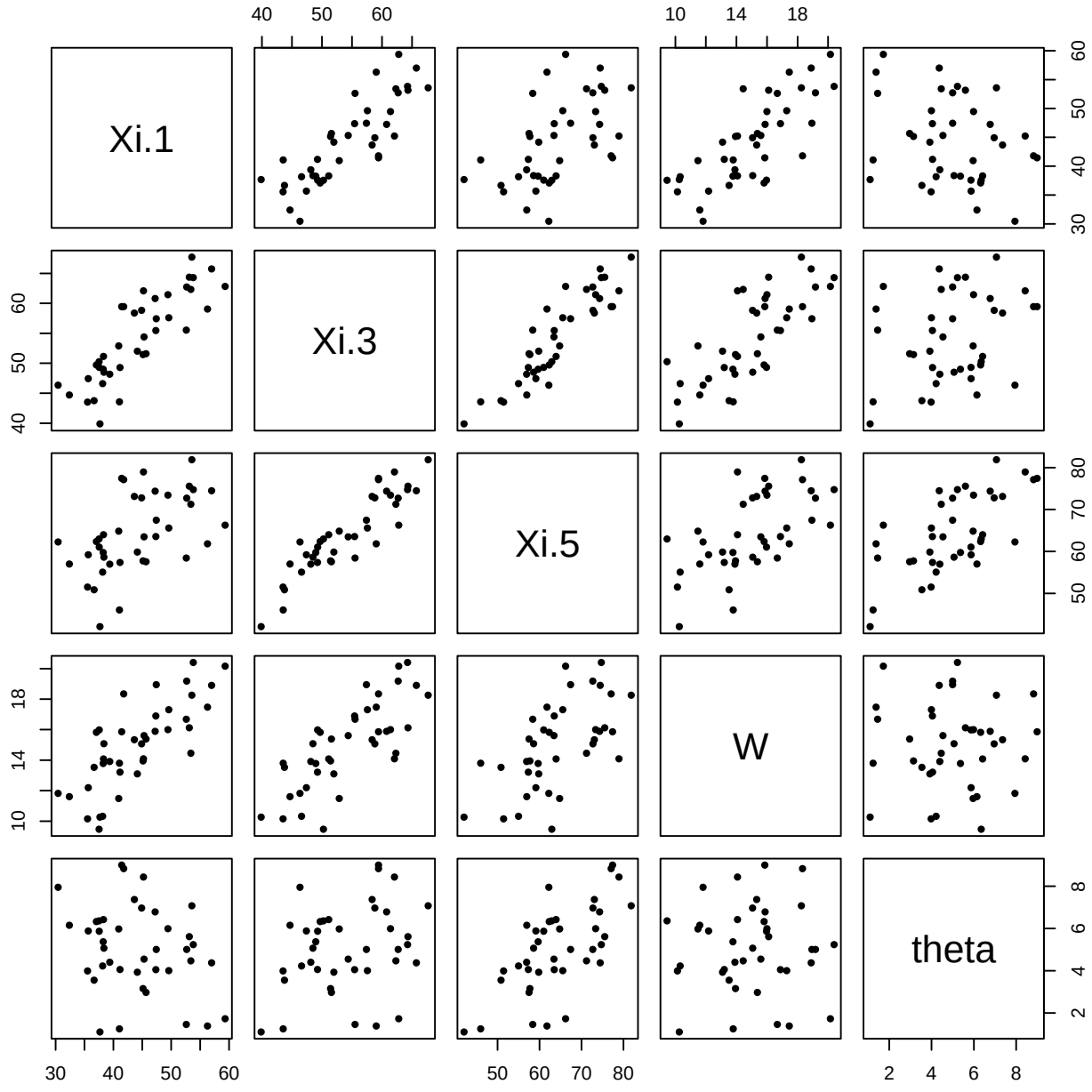
Data for Exhibit 1

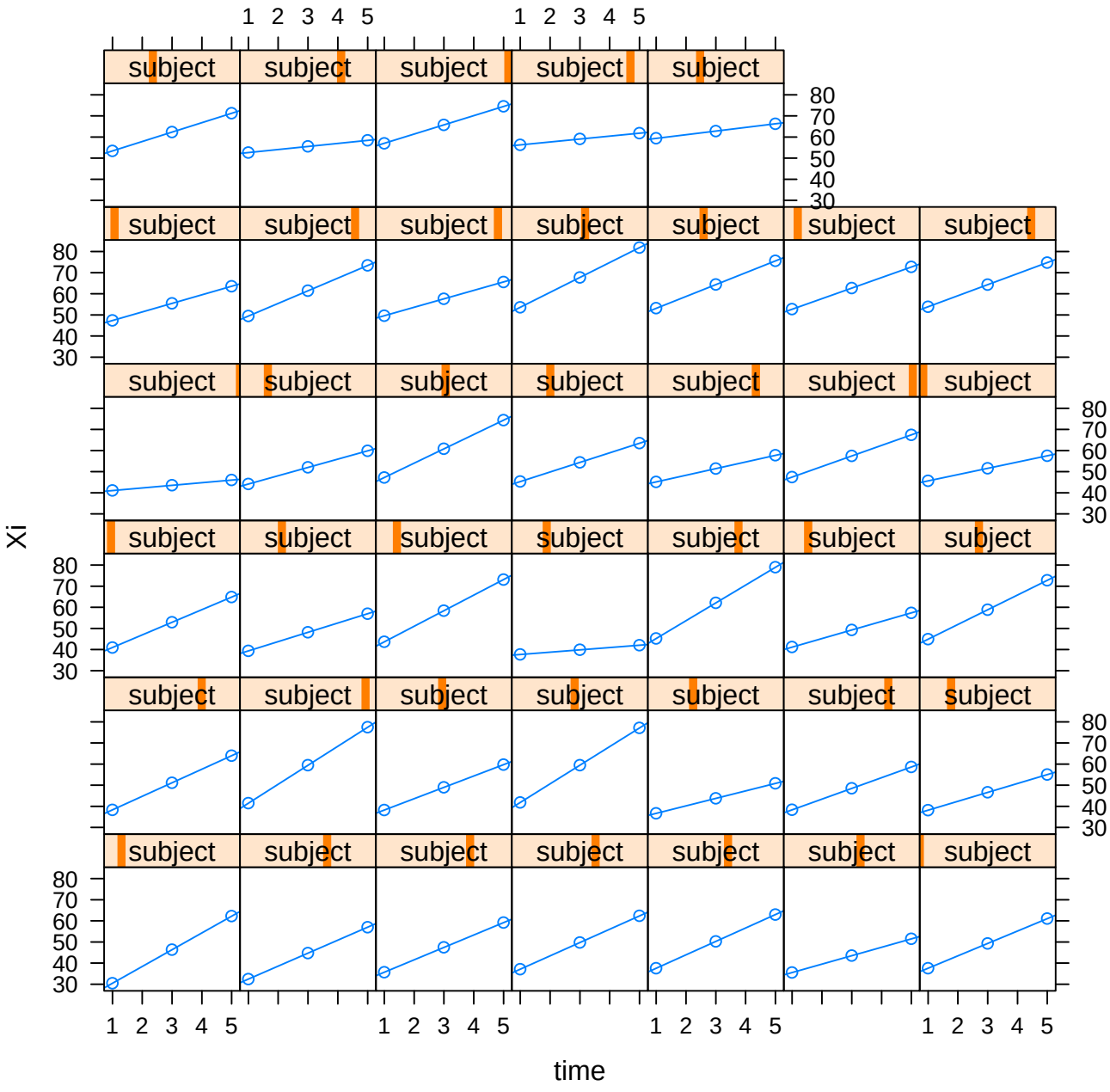
	Xi(1)	Xi(3)	Xi(5)	W
1	37.559130	49.290530	61.021930	15.972470
2	45.654290	51.584510	57.514720	15.377240
3	40.938810	52.879780	64.820760	11.479020
4	47.359370	55.448790	63.538220	16.889440
5	52.705110	62.703510	72.701910	19.178340
6	30.452310	46.340820	62.229340	11.818220
7	43.646250	58.370030	73.093820	15.328750
8	41.155490	49.262760	57.370030	13.208130
9	44.151480	51.998020	59.844550	13.090430
10	38.159650	46.594290	55.028920	10.315590
11	37.675940	39.867320	42.058700	10.261310
12	45.300540	54.382830	63.465110	15.598520
13	39.369470	48.153540	56.937610	13.900920
14	36.663710	43.751210	50.838700	13.525720
15	53.398540	62.316440	71.234350	14.447020
16	59.354590	62.802520	66.250450	20.158750
17	53.139720	64.349090	75.558460	16.114490
18	44.901730	58.824930	72.748130	15.057730
19	41.786250	59.440230	77.094210	18.333810
20	38.245640	48.980320	59.714990	13.772200
21	47.235960	60.788930	74.341890	15.882300
22	53.571270	67.708060	81.844860	18.253550
23	35.542900	43.510950	51.479000	10.145410
24	37.543520	50.248820	62.954120	9.461730
25	37.065520	49.707010	62.348510	15.814920
26	32.398090	44.689060	56.980030	11.604630
27	45.216440	62.076580	78.936720	14.077550
28	35.671760	47.421170	59.170580	12.186710
29	38.301750	51.134650	63.967540	14.072240
30	52.613470	55.517540	58.421610	16.679830
31	38.362050	48.490300	58.618560	15.071560
32	45.139850	51.435610	57.731370	13.942930
33	53.819050	64.274460	74.729870	20.399220
34	49.455840	61.424760	73.393680	15.996710
35	56.285520	59.042180	61.798830	17.467350
36	49.588300	57.577850	65.567410	17.296230
37	41.448820	59.431220	77.413640	15.857430
38	47.417680	57.421590	67.425480	18.946520
39	56.998030	65.732350	74.466660	18.896400
40	41.060790	43.543620	46.026450	13.790020

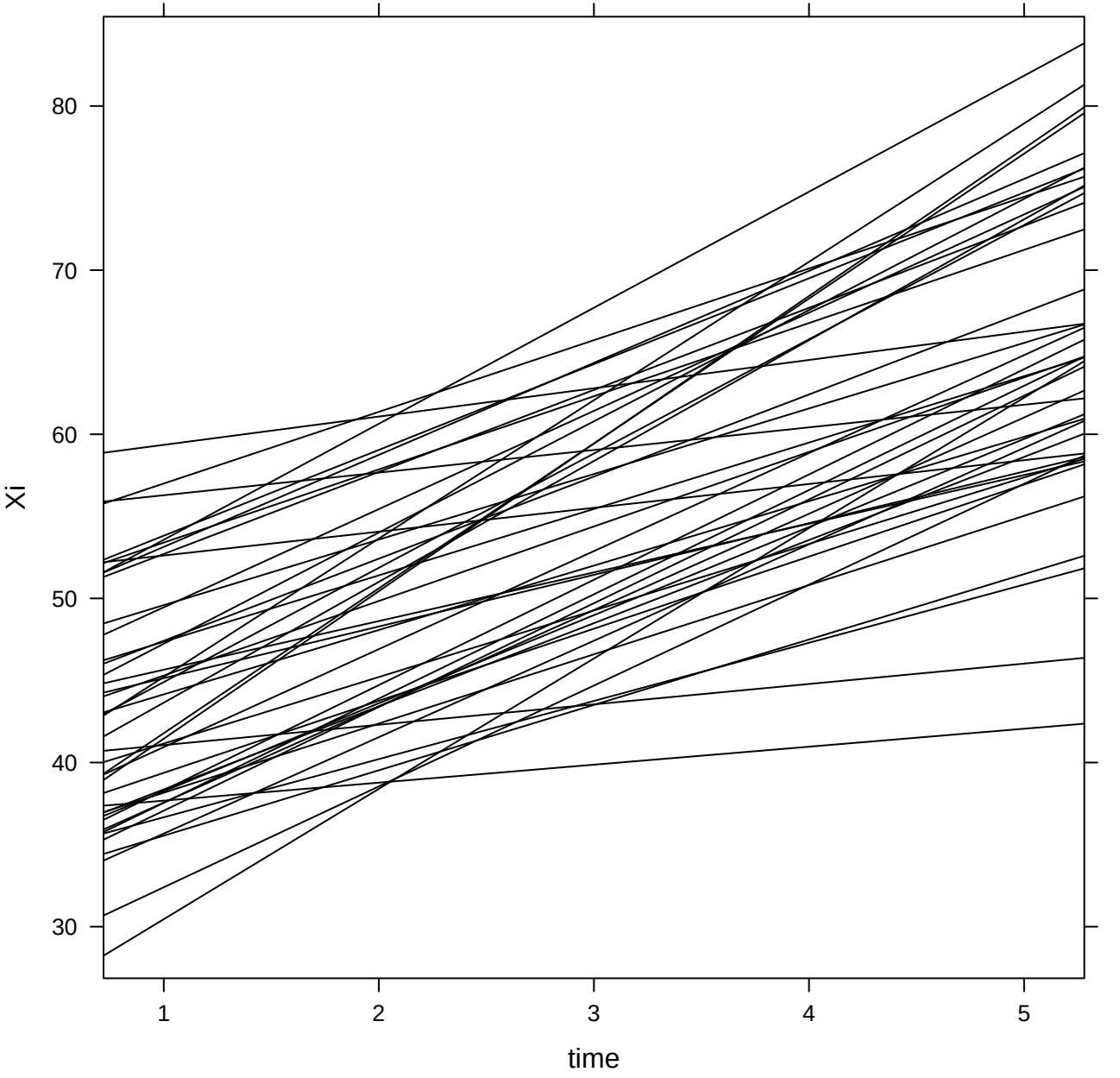
==========

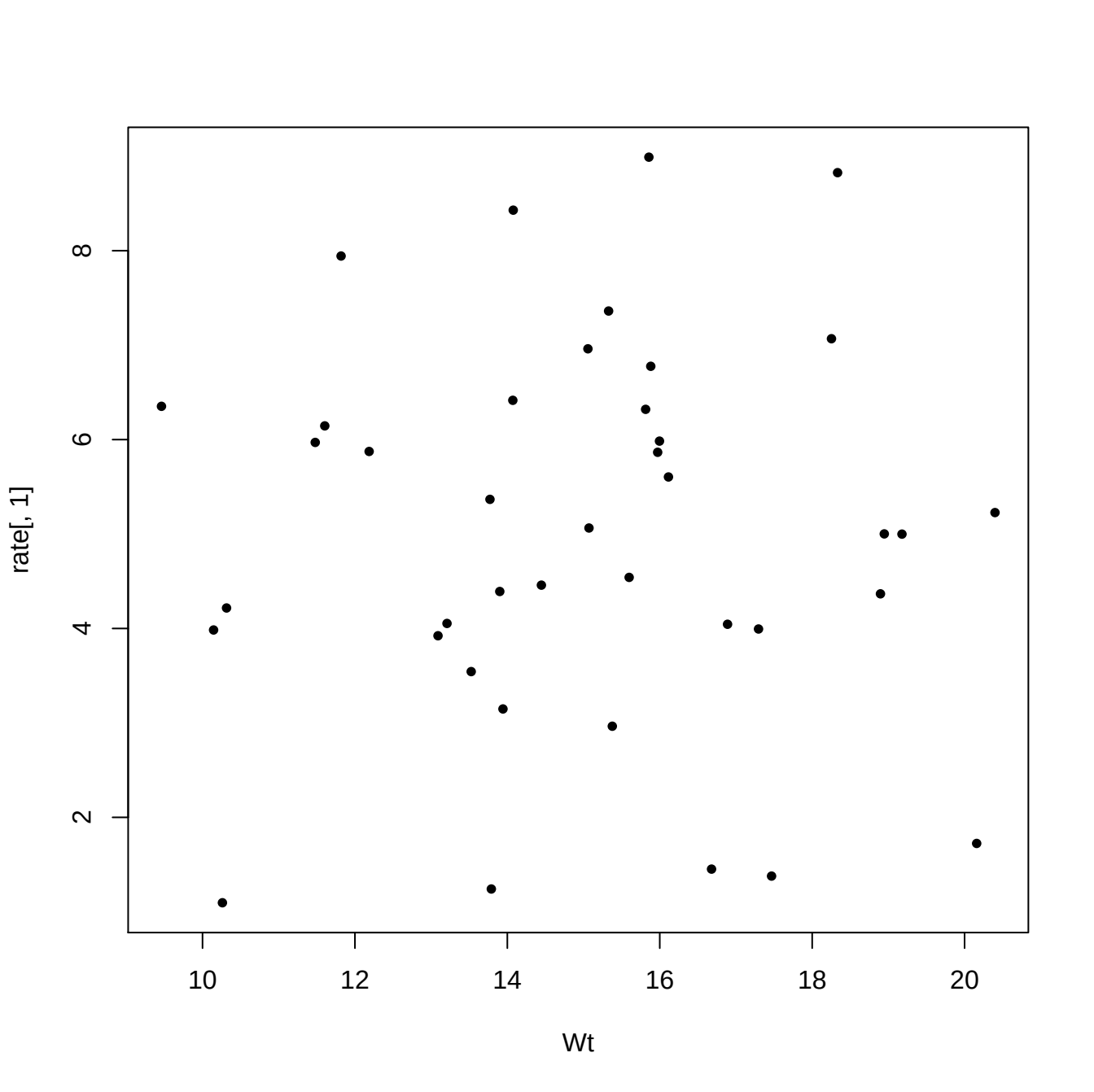
Corresponding Data for Fallible Observations (X)

	X(1)	X(3)	X(5)	W
1	37.516320	51.352380	59.447650	15.972470
2	45.127490	52.817920	61.646580	15.377240
3	35.146190	56.825750	66.150560	11.479020
4	44.125920	49.189990	64.570750	16.889440
5	52.742550	66.558240	70.488200	19.178340
6	30.429370	49.953630	64.290860	11.818220
7	45.855950	61.804990	68.040070	15.328750
8	41.085170	48.477920	56.037560	13.208130
9	45.596330	53.609550	56.391610	13.090430
10	41.640850	52.921170	53.426490	10.315590
11	40.553350	41.063000	42.669360	10.261310
12	43.596080	50.701220	61.301810	15.598520
13	40.330890	42.926600	56.823440	13.900920
14	36.468130	39.048250	55.981900	13.525720
15	50.935130	64.577550	73.268780	14.447020
16	56.389270	64.351740	66.465300	20.158750
17	54.820850	55.940590	78.981190	16.114490
18	46.234100	55.570820	69.208680	15.057730
19	40.338170	55.815210	79.839300	18.333810
20	39.782590	48.463580	61.588510	13.772200
21	45.568470	57.297970	76.471220	15.882300
22	50.794400	66.029400	82.140590	18.253550
23	36.556680	45.836010	41.518490	10.145410
24	39.484500	50.684350	57.499870	9.461730

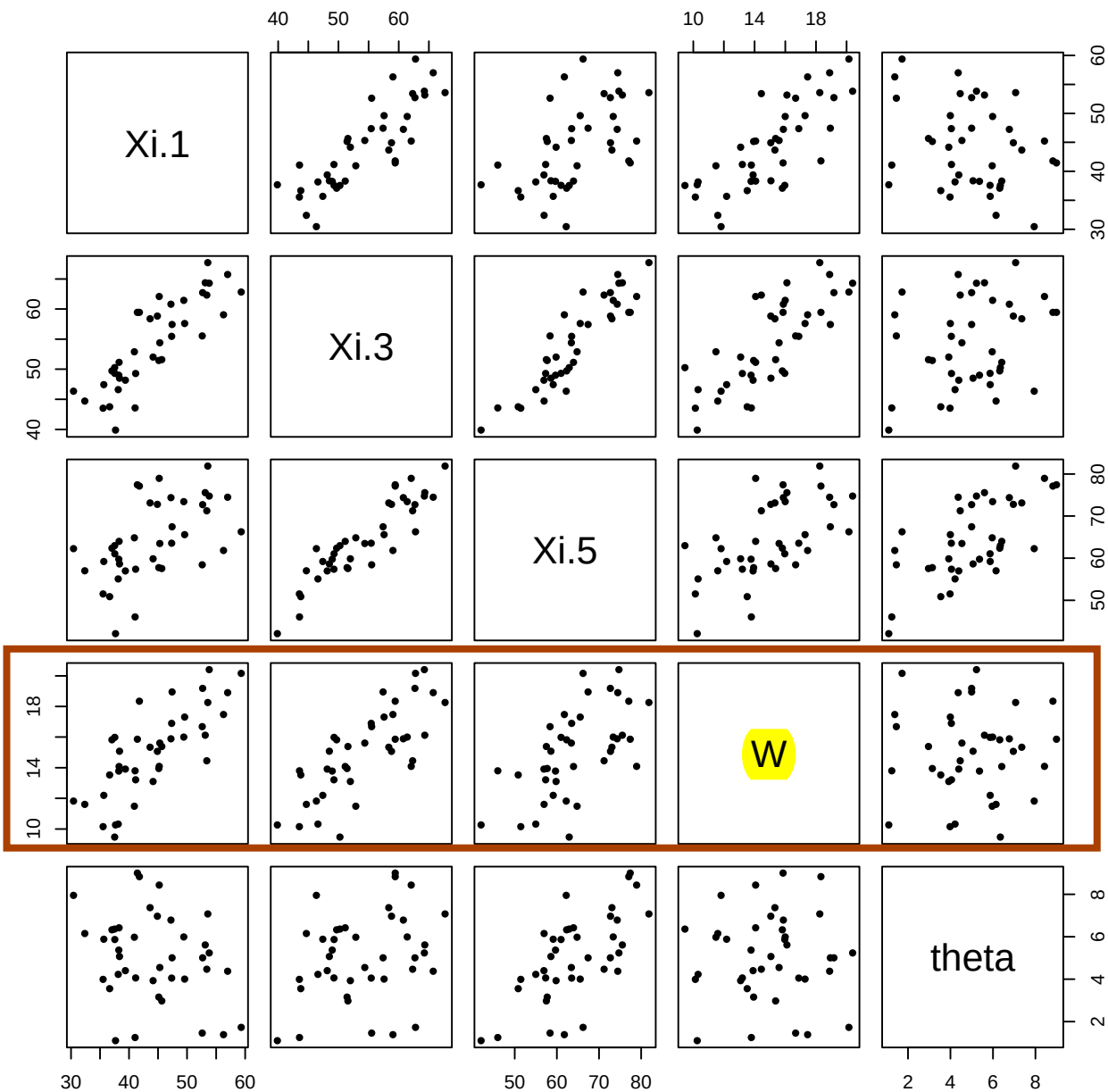

Xi.1
Xi.3
Xi.5
W
theta
40
50
60
10
14
18
30
40
50
60
50
60
70
80
2
4
6
8

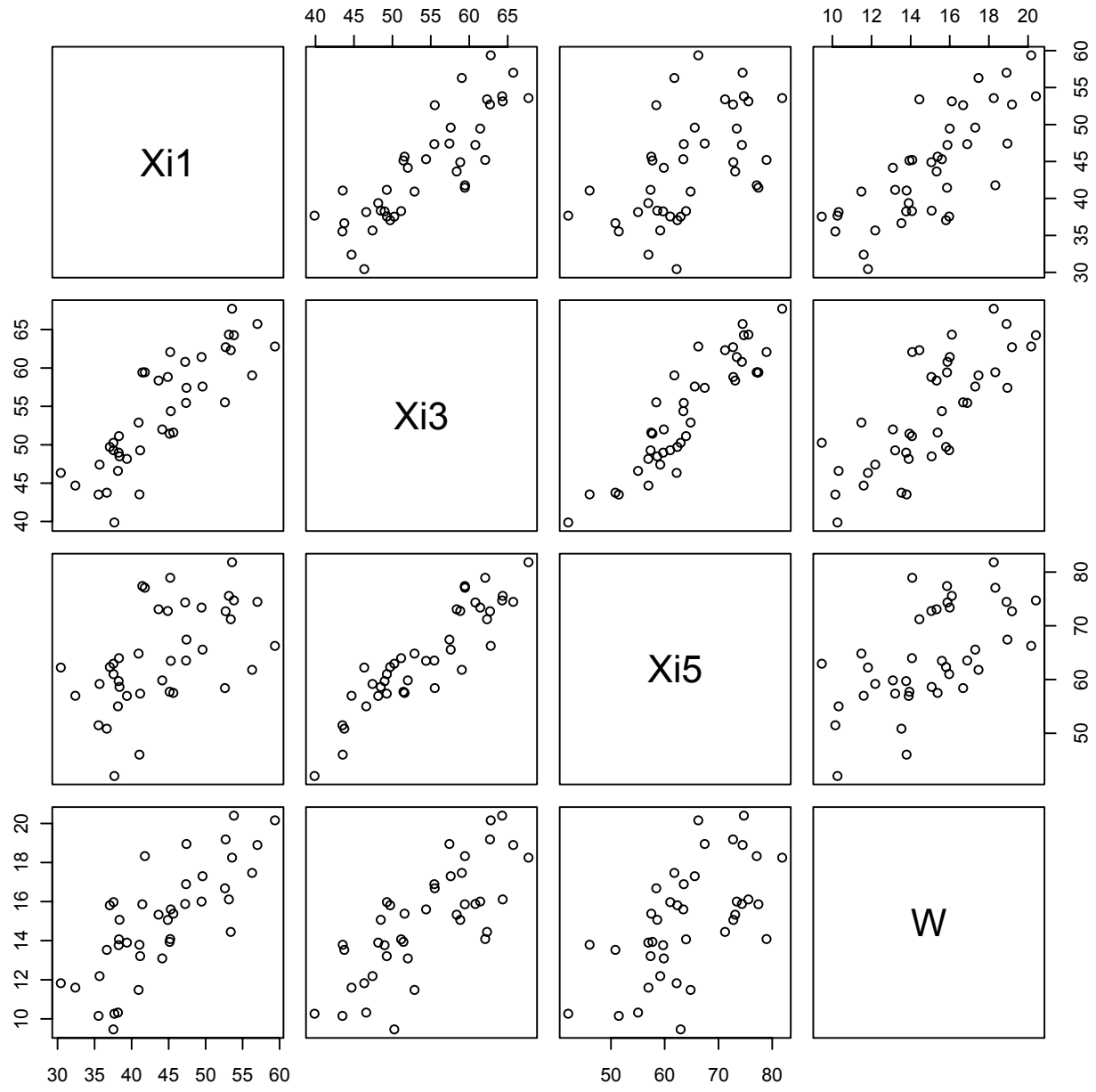

subject
Xi
time
1 2 3 4 5
30 40 50 60 70 80


80
70
60
50
40
30
Xi
1
2
3
4
5
time


rate[, 1]
2
4
6
8
10
12
14
16
18
20
Wt

Myths Data




Xi1
Xi3
Xi5
W
40 45 50 55 60 65
10 12 14 16 18 20
30 35 40 45 50 55 60
50 60 70 80

DRogosa STAT 222 week 4

measurement of change examples

```
Time1-Time2 regressions
Example from  Rogosa, D. R. (1995). Myths and methods: "Myths about longitudinal
research," plus supplemental questions. In The analysis of change, J. M.
Gottman, Ed. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 3-65.

> mtruesig$theta = signif((mtrue$Xi5 - mtrue$Xi1)/4,4)
> mtruesig
     Xi1   Xi3   Xi5       W theta
1  37.56 49.29 61.02 15.970 5.866
2  45.65 51.58 57.51 15.380 2.965
3  40.94 52.88 64.82 11.480 5.970
4  47.36 55.45 63.54 16.890 4.045
5  52.71 62.70 72.70 19.180 4.999
6  30.45 46.34 62.23 11.820 7.944
7  43.65 58.37 73.09 15.330 7.362
8  41.16 49.26 57.37 13.210 4.054
9  44.15 52.00 59.84 13.090 3.923
10 38.16 46.59 55.03 10.320 4.217
11 37.68 39.87 42.06 10.260 1.096
12 45.30 54.38 63.47 15.600 4.541
13 39.37 48.15 56.94 13.900 4.392
14 36.66 43.75 50.84 13.530 3.544
15 53.40 62.32 71.23 14.450 4.459
16 59.35 62.80 66.25 20.160 1.724
17 53.14 64.35 75.56 16.110 5.605
18 44.90 58.82 72.75 15.060 6.962
19 41.79 59.44 77.09 18.330 8.827
20 38.25 48.98 59.71 13.770 5.367
21 47.24 60.79 74.34 15.880 6.776
22 53.57 67.71 81.84 18.250 7.068
23 35.54 43.51 51.48 10.150 3.984
24 37.54 50.25 62.95  9.462 6.353
25 37.07 49.71 62.35 15.810 6.321
26 32.40 44.69 56.98 11.600 6.145
27 45.22 62.08 78.94 14.080 8.430
28 35.67 47.42 59.17 12.190 5.875
29 38.30 51.13 63.97 14.070 6.416
30 52.61 55.52 58.42 16.680 1.452
31 38.36 48.49 58.62 15.070 5.064
32 45.14 51.44 57.73 13.940 3.148
33 53.82 64.27 74.73 20.400 5.228
34 49.46 61.42 73.39 16.000 5.984
35 56.29 59.04 61.80 17.470 1.378
36 49.59 57.58 65.57 17.300 3.995
37 41.45 59.43 77.41 15.860 8.991
38 47.42 57.42 67.43 18.950 5.002
39 57.00 65.73 74.47 18.900 4.367
40 41.06 43.54 46.03 13.790 1.241
> pairs(~ Xi1 + Xi3 + Xi5 + W)
> cor(mtruesig)
            Xi1       Xi3       Xi5            W        theta
Xi1   1.0000000 0.8422138 0.5359036  0.766175758 -0.280851506
Xi3   0.8422138 1.0000000 0.9065331  0.765188951  0.280906648
Xi5   0.5359036 0.9065331 1.0000000  0.598501096  0.659788513
W     0.7661758 0.7651890 0.5985011  1.000000000 -0.001592367
theta -0.2808515 0.2809066 0.6597885 -0.001592367  1.000000000
```

link on class page

Artificial Data
perfect meas
exact straight line
$\mu_\theta = 5$

look real see pdf.

rate of change (theta)

```
> truereg1 = lm(Xi5 ~ W + Xi1)
> truereg2 = lm(Xi5 ~ W + Xi3)
> truediffreg = lm(I(Xi5- Xi3) ~ W)
> summary(truereg1)
Call: lm(formula = Xi5 ~ W + Xi1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  31.2139     7.5445   4.137 0.000194 ***
W             1.5002     0.6680   2.246 0.030788 *
Xi1           0.2392     0.2588   0.924 0.361290
---
Residual standard error: 7.514 on 37 degrees of freedom
Multiple R-squared: 0.3727,     Adjusted R-squared: 0.33
F-statistic: 10.99 on 2 and 37 DF,  p-value: 0.0001792
```

perfect meas. W important predictor of change

```
> summary(truereg2)
Call: lm(formula = Xi5 ~ W + Xi3)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6874     4.5537   0.151   0.8808
W            -0.7570     0.3329  -2.274   0.0289 *
Xi3           1.3821     0.1290  10.718   6.7e-13 ***
---
Residual standard error: 3.751 on 37 degrees of freedom
Multiple R-squared: 0.8437,     Adjusted R-squared: 0.83
F-statistic: 99.83 on 2 and 37 DF,  p-value: 1.232e-15
```

perfect meas W important negative predictor of change

```
> summary(truediffreg)
Call: lm(formula = I(Xi5 - Xi3) ~ W)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.086567   3.586436   2.812  0.00774 **
W           -0.002139   0.235245  -0.009  0.99279
---
Residual standard error: 4.117 on 38 degrees of freedom
Multiple R-squared: 2.176e-06,  Adjusted R-squared: -0.0
F-statistic: 8.267e-05 on 1 and 38 DF,  p-value: 0.9928

> cor(W, theta)   [1] -0.001592367
```

change Δ on W no relation

create straight-line data W s.t. $\rho_{\theta W} = 0$
try out standard regression approaches using prior X_is as predictor

```
##Same result if Difference Score is Outcome rather than final status

#First the true score regressions from [redacted] handout
> truereg1D = lm(I(Xi5 - Xi1) ~ W + Xi1)
> summary(truereg1D)

Call:
lm(formula = I(Xi5 - Xi1) ~ W + Xi1)

Residuals:
    Min      1Q  Median      3Q     Max
-15.692  -4.348  -1.051   6.406  15.788

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  31.2139     7.5445   4.137 0.000194 ***
W             1.5002     0.6680   2.246 0.030788 *
Xi1          -0.7608     0.2588  -2.940 0.005624 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.
Residual standard error: 7.514 on 37 degrees of freedom
Multiple R-squared: 0.1894,     Adjusted R-squared: 0.1
F-statistic: 4.323 on 2 and 37 DF,  p-value: 0.02055
```

match coeff, t-statistic

```
> truereg2D = lm(I(Xi5 - Xi3) ~ W + Xi3)
> summary(truereg2D)

Call:
lm(formula = I(Xi5 - Xi3) ~ W + Xi3)

Residuals:
     Min       1Q   Median       3Q      Max
-7.26371 -2.36848 -0.07474  2.20751  8.12447

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6874     4.5537   0.151  0.88083
W            -0.7570     0.3329  -2.274  0.02886 *
Xi3           0.3821     0.1290   2.963  0.00529 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.751 on 37 degrees of freedom
Multiple R-squared: 0.1918,     Adjusted R-squared: 0.1481
F-statistic: 4.391 on 2 and 37 DF,  p-value: 0.01945

> detach(mtruesig)
```

match coeff t-stat

```
Time1-Time2 regressions
Example from   Rogosa, D. R. (1995). Myths and methods: "Myths about longitudinal
research," plus supplemental questions. In The analysis of change, J. M.
Gottman, Ed. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 3-65.

> mtruesig$theta = signif((mtrue$Xi5 - mtrue$Xi1)/4,4)
> mtruesig
     Xi1   Xi3   Xi5      W theta
1  37.56 49.29 61.02 15.970 5.866
2  45.65 51.58 57.51 15.380 2.965
3  40.94 52.88 64.82 11.480 5.970
4  47.36 55.45 63.54 16.890 4.045
5  52.71 62.70 72.70 19.180 4.999
6  30.45 46.34 62.23 11.820 7.944
7  43.65 58.37 73.09 15.330 7.362
8  41.16 49.26 57.37 13.210 4.054
9  44.15 52.00 59.84 13.090 3.923
10 38.16 46.59 55.03 10.320 4.217
11 37.68 39.87 42.06 10.260 1.096
12 45.30 54.38 63.47 15.600 4.541
13 39.37 48.15 56.94 13.900 4.392
14 36.66 43.75 50.84 13.530 3.544
15 53.40 62.32 71.23 14.450 4.459
16 59.35 62.80 66.25 20.160 1.724
17 53.14 64.35 75.56 16.110 5.605
18 44.90 58.82 72.75 15.060 6.962
19 41.79 59.44 77.09 18.330 8.827
20 38.25 48.98 59.71 13.770 5.367
21 47.24 60.79 74.34 15.880 6.776
22 53.57 67.71 81.84 18.250 7.068
23 35.54 43.51 51.48 10.150 3.984
24 37.54 50.25 62.95  9.462 6.353
25 37.07 49.71 62.35 15.810 6.321
26 32.40 44.69 56.98 11.600 6.145
27 45.22 62.08 78.94 14.080 8.430
28 35.67 47.42 59.17 12.190 5.875
29 38.30 51.13 63.97 14.070 6.416
30 52.61 55.52 58.42 16.680 1.452
31 38.36 48.49 58.62 15.070 5.064
32 45.14 51.44 57.73 13.940 3.148
33 53.82 64.27 74.73 20.400 5.228
34 49.46 61.42 73.39 16.000 5.984
35 56.29 59.04 61.80 17.470 1.378
36 49.59 57.58 65.57 17.300 3.995
37 41.45 59.43 77.41 15.860 8.991
38 47.42 57.42 67.43 18.950 5.002
39 57.00 65.73 74.47 18.900 4.367
40 41.06 43.54 46.03 13.790 1.241
> pairs(~ Xi1 + Xi3 + Xi5 + W)
> cor(mtruesig)
             Xi1       Xi3       Xi5           W        theta
Xi1    1.0000000 0.8422138 0.5359036  0.766175758 -0.280851506
Xi3    0.8422138 1.0000000 0.9065331  0.765188951  0.280906648
Xi5    0.5359036 0.9065331 1.0000000  0.598501096  0.659788513
W      0.7661758 0.7651890 0.5985011  1.000000000 -0.001592367
theta -0.2808515 0.2809066 0.6597885 -0.001592367  1.000000000
```

```
> truereg1 = lm(Xi5 ~ W + Xi1)
> truereg2 = lm(Xi5 ~ W + Xi3)
> truediffreg = lm(I(Xi5- Xi3) ~ W)
> summary(truereg1)
Call: lm(formula = Xi5 ~ W + Xi1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  31.2139     7.5445   4.137 0.000194 ***
W             1.5002     0.6680   2.246 0.030788 *
Xi1           0.2392     0.2588   0.924 0.361290
---
Residual standard error: 7.514 on 37 degrees of freedom
Multiple R-squared: 0.3727,     Adjusted R-squared: 0.33
F-statistic: 10.99 on 2 and 37 DF,  p-value: 0.0001792

> summary(truereg2)
Call: lm(formula = Xi5 ~ W + Xi3)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6874     4.5537   0.151   0.8808
W            -0.7570     0.3329  -2.274   0.0289 *
Xi3           1.3821     0.1290  10.718  6.7e-13 ***
---
Residual standard error: 3.751 on 37 degrees of freedom
Multiple R-squared: 0.8437,     Adjusted R-squared: 0.83
F-statistic: 99.83 on 2 and 37 DF,  p-value: 1.232e-15

> summary(truediffreg)
Call: lm(formula = I(Xi5 - Xi3) ~ W)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.086567   3.586436   2.812  0.00774 **
W           -0.002139   0.235245  -0.009  0.99279
---
Residual standard error: 4.117 on 38 degrees of freedom
Multiple R-squared: 2.176e-06,  Adjusted R-squared: -0.0
F-statistic: 8.267e-05 on 1 and 38 DF,  p-value: 0.9928

> cor(W, theta)   [1] -0.001592367
```

continued

```
#First the true score regressions from class 3/3 handout
> truereg1D = lm(I(Xi5 - Xi1) ~ W + Xi1)
> summary(truereg1D)

Call:
lm(formula = I(Xi5 - Xi1) ~ W + Xi1)

Residuals:
    Min      1Q  Median      3Q     Max
-15.692  -4.348  -1.051   6.406  15.788

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  31.2139     7.5445   4.137 0.000194 ***
W             1.5002     0.6680   2.246 0.030788 *
Xi1          -0.7608     0.2588  -2.940 0.005624 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.514 on 37 degrees of freedom
Multiple R-squared: 0.1894,     Adjusted R-squared: 0.1456
F-statistic: 4.323 on 2 and 37 DF,  p-value: 0.02055

> truereg2D = lm(I(Xi5 - Xi3) ~ W + Xi3)
> summary(truereg2D)

Call:
lm(formula = I(Xi5 - Xi3) ~ W + Xi3)

Residuals:
     Min       1Q   Median       3Q      Max
-7.26371 -2.36848 -0.07474  2.20751  8.12447

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6874     4.5537   0.151  0.88083
W            -0.7570     0.3329  -2.274  0.02886 *
Xi3           0.3821     0.1290   2.963  0.00529 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.751 on 37 degrees of freedom
Multiple R-squared: 0.1918,     Adjusted R-squared: 0.1481
F-statistic: 4.391 on 2 and 37 DF,  p value: 0.01945

> detach(mtruesig)
```

match coeff, t-stat

match coeff, t-stat

STAT 222
week 1

Demonstration:
time 1, time 2 Regressions

DATA Y_2, Y_1, W exog.

D on Y_1, W same as Y_2 on Y_1, W

coefficients for W in population or sample,

perfect or fallible measurement

$$\text{coeff } \beta_{DW \cdot Y_1} = \beta_{D(W \cdot Y_1)} = \frac{Cov(Y_2 - Y_1, W - \beta_{WY_1} Y_1)}{Var(W \cdot Y_1)}$$

(Note: $\beta_{WY_1} Var(Y_1) = Cov(Y_1, W)$)

$$= \frac{Cov(Y_2, W) - \beta_{WY_1} Cov(Y_1, Y_2) - Cov(Y_1, W) + \beta_{WY_1} Var(Y_1)}{Var(W \cdot Y_1)}$$

$$= \frac{Cov(Y_2, W) - \beta_{WY_1} Cov(Y_1, Y_2)}{Var(W \cdot Y_1)}$$

$$= \frac{Cov(Y_2, W \cdot Y_1)}{Var(W \cdot Y_1)} = \beta_{Y_2(W \cdot Y_1)} = \beta_{Y_2 W \cdot Y_1}$$

brute force, quicker ways to this

see STAT 209 week 1 adjusted variables

```
    #Stat222, Week 1 example, Rogosa R-session  4/8/12

R version 2.14.1 (2011-12-22)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-pc-mingw32/x64 (64-bit)


> week1Xi = read.table(file="D:\\drr12\\stat222\\week1\\mythdata_Xi", header = T)
#I took the web page and commented out via "#" all the lines except the Xi-data (40 row
# I named the observation columns as shown below

> head(week1Xi)
      Xi.1     Xi.3     Xi.5        W
1 37.55913 49.29053 61.02193 15.97247
2 45.65429 51.58451 57.51472 15.37724
3 40.93881 52.87978 64.82076 11.47902
4 47.35937 55.44879 63.53822 16.88944
5 52.70511 62.70351 72.70191 19.17834
6 30.45231 46.34082 62.22934 11.81822
> week1Xi$theta = (week1Xi$Xi.5 - week1Xi$Xi.1)/4 # create the "theta" column in the we
# this works only because the "Xi" data fall exactly on a straight-line (illustrated be
> head(week1Xi)
      Xi.1     Xi.3     Xi.5        W    theta
1 37.55913 49.29053 61.02193 15.97247 5.865700
2 45.65429 51.58451 57.51472 15.37724 2.965107
3 40.93881 52.87978 64.82076 11.47902 5.970488
4 47.35937 55.44879 63.53822 16.88944 4.044713
5 52.70511 62.70351 72.70191 19.17834 4.999200
6 30.45231 46.34082 62.22934 11.81822 7.944257
> attach(week1Xi)

> cor(W,theta)
[1] -0.001411346

> cor(week1Xi)
            Xi.1      Xi.3      Xi.5            W        theta
Xi.1   1.0000000 0.8421714 0.5357932  0.765952711 -0.280944258
Xi.3   0.8421714 1.0000000 0.9065112  0.765172576  0.280889494
Xi.5   0.5357932 0.9065112 1.0000000  0.598471157  0.659814293
W      0.7659527 0.7651726 0.5984712  1.000000000 -0.001411346
theta -0.2809443 0.2808895 0.6598143 -0.001411346  1.000000000
> pairs(week1Xi)
> pairs(week1Xi, pch = 20)  # this is the plot that is posted in the plot/link

> #do the regressions from the week 1 handout
> truereg1 = lm(Xi.5 ~ W + Xi.1)
> truereg2 = lm(Xi.5 ~ W + Xi.3)
> truereg3 = lm(theta ~ W )
> summary(truereg1)

Call:
lm(formula = Xi.5 ~ W + Xi.1)

Residuals:
    Min      1Q  Median      3Q     Max
-15.697  -4.351  -1.048   6.413  15.788

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  31.2134     7.5457   4.137 0.000195 ***
W             1.5004     0.6678   2.247 0.030712 *
Xi.1          0.2392     0.2587   0.925 0.361216
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.514 on 37 degrees of freedom
Multiple R-squared: 0.3727,     Adjusted R-squared: 0.3387
F-statistic: 10.99 on 2 and 37 DF,  p-value: 0.0001794

> summary(truereg2)

Call:
lm(formula = Xi.5 ~ W + Xi.3)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2692 -2.3773 -0.0794  2.2062  8.1319

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6830     4.5547   0.150   0.8816
W            -0.7570     0.3329  -2.274   0.0289 *
Xi.3          1.3822     0.1290  10.717 6.72e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.752 on 37 degrees of freedom
Multiple R-squared: 0.8436,     Adjusted R-squared: 0.8352
F-statistic:  99.8 on 2 and 37 DF,  p-value: 1.238e-15

> summary(truereg3)

Call:
lm(formula = theta ~ W)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9362 -1.0344  0.0081  1.3009  3.9650

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.042388   1.793449   2.812  0.00776 **
W           -0.001023   0.117641  -0.009  0.99310
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.059 on 38 degrees of freedom
Multiple R-squared: 1.992e-06,  Adjusted R-squared: -0.02631
F-statistic: 7.569e-05 on 1 and 38 DF,  p-value: 0.9931

> confint(truereg1) #I'm sure you did this sort of thing in your intro courses
                 2.5 %     97.5 %
(Intercept) 15.9243934 46.5023390
W            0.1472643  2.8534413
Xi.1        -0.2849891  0.7632945
> confint(truereg2)
                2.5 %      97.5 %
(Intercept) -8.545741  9.91166842
W           -1.431584 -0.08237203
```

```
Xi.3         1.120864  1.64351319
> confint(truereg3)
                2.5 %    97.5 %
(Intercept)  1.4117397 8.6730369
W           -0.2391757 0.2371287
> truereg1a = lm(theta ~ W + Xi.1)
> truereg2a = lm(theta ~ W + Xi.3)
> summary(truereg1a) # do the lower frame examples with change as the outcome
  # because I used rate rather than amount of change to match coeefs you need to scale

Call:
lm(formula = theta ~ W + Xi.1)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9242 -1.0879 -0.2619  1.6032  3.9471

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.80334    1.88642   4.137 0.000195 ***
W            0.37509    0.16695   2.247 0.030712 *
Xi.1        -0.19021    0.06467  -2.941 0.005610 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.879 on 37 degrees of freedom
Multiple R-squared: 0.1895,     Adjusted R-squared: 0.1457
F-statistic: 4.325 on 2 and 37 DF,  p-value: 0.02051

> summary(truereg2a)

Call:
lm(formula = theta ~ W + Xi.3)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6346 -1.1886 -0.0397  1.1031  4.0660

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.34148    2.27735   0.150   0.8816
W           -0.37849    0.16647  -2.274   0.0289 *
Xi.3         0.19109    0.06449   2.963   0.0053 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.876 on 37 degrees of freedom
Multiple R-squared: 0.1918,     Adjusted R-squared: 0.1481
F-statistic: 4.391 on 2 and 37 DF,  p-value: 0.01945

> confint(truereg1a)
                2.5 %      97.5 %
(Intercept)  3.98109835 11.62558475
W            0.03681608  0.71336033
Xi.1        -0.32124729 -0.05917636
> confint(truereg2a)
                2.5 %      97.5 %
(Intercept) -4.27286705  4.95583457
W           -0.71579184 -0.04118593
Xi.3         0.06043203  0.32175647
```

COLLOQUIUM PAPER

Metaanalysis of the relationship between violent video game play and physical aggression over time

Anna T. Prescott[a], James D. Sargent[b], and Jay G. Hull[a,1]

[a]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH 03755; and [b]Department of Pediatrics, Geisel School of Medicine, Dartmouth College, Hanover, NH 03755

Edited by David E. Meyer, University of Michigan, Ann Arbor, MI, and approved August 10, 2017 (received for review August 27, 2016)

To clarify and quantify the influence of video game violence (VGV) on aggressive behavior, we conducted a metaanalysis of all prospective studies to date that assessed the relation between exposure to VGV and subsequent overt physical aggression. The search strategy identified 24 studies with over 17,000 participants and time lags ranging from 3 months to 4 years. The samples comprised various nationalities and ethnicities with mean ages from 9 to 19 years. For each study we obtained the standardized regression coefficient for the prospective effect of VGV on subsequent aggression, controlling for baseline aggression. VGV was related to aggression using both fixed [β = 0.113, 95% CI = (0.098, 0.128)] and random effects models [β = 0.106 (0.078, 0.134)]. When all available covariates were included, the size of the effect remained significant for both models [β = 0.080 (0.065, 0.094) and β = 0.078 (0.053, 0.102), respectively]. No evidence of publication bias was found. Ethnicity was a statistically significant moderator for the fixed-effects models ($P \leq 0.011$) but not for the random-effects models. Stratified analyses indicated the effect was largest among Whites, intermediate among Asians, and nonsignificant among Hispanics. Discussion focuses on the implications of such findings for current debates regarding the effects of violent video games on physical aggression.

video games | aggression | metaanalysis | ethnicity | longitudinal

A controversy has developed over the relation of violent video game play and aggression (1–4). Whereas the majority of those who conduct research on this topic argue that playing such games increases aggressive behavior, a vocal minority has argued that the relation of game play and real-world aggressive behavior is at best overstated and at worst spurious. The controversy has had important real-world implications. In 2011, the US Supreme Court struck down a California statute designed to limit purchases and rentals of extremely violent video games by children (5). The majority opinion expressed skepticism about the importance of effects of violent video games, likening them to a "harmless pastime" (5).

Violent Video Game Play and Aggression

The case that violent video game play increases aggressive behavior has been made most forcefully by Anderson et al. (6; see also refs. 7 and 8). Specifically, these authors undertook a comprehensive metaanalysis of the literature on the impact of violent video game play on six categories of aggressive response: cognition, affect, arousal, empathy/sensitization to violence, overt aggressive behavior, and overt prosocial behavior. Their metaanalysis examined effects from over 130 research reports based on over 130,000 participants. On the basis of these analyses, the authors concluded that violent video game play is positively associated with aggressive behavior, aggressive cognition, and aggressive affect, as well as negatively associated with empathy for victims of violence and with prosocial behavior. Furthermore, the authors concluded that these effects are statistically reliable in experimental, cross-sectional, and longitudinal studies, are observed across cultures, gender, and game types (e.g., first vs. third person perspective; human vs. nonhuman targets; and so forth), and that methodologically superior studies tended to yield larger effects. A more recent metaanalysis by Greitemeyer and Mügge (9) came to similar conclusions.

Although hailed by some as conclusively demonstrating a link between violent video game play and aggression (7), the Anderson et al. (6) metaanalysis did not decrease skepticism among a vocal minority of researchers (10). In a wide range of articles, Ferguson (2, 11–16) has leveled four criticisms at research purporting to show that video game violence (VGV) increases real-world aggression: (*i*) many studies that support such a link use measures of "nonserious aggression" (e.g., accessibility of aggression related words, aggression related feelings) that inflate effect-size estimates; (*ii*) many studies do not include important covariates as statistical controls and hence any observed effects may be spurious consequences of third variable relationships; (*iii*) there is a bias to publish studies supporting a VGV → aggression link as opposed to those reporting a null effect; and (*iv*) even if one accepts the existence of a VGV → aggression relationship, the estimated effect size typically reported is exceedingly weak. Despite the fact that these arguments have been vigorously rebutted by Anderson and his colleagues (8), Ferguson and his colleagues have continued to stand by their critique (2, 15, 17, 18). With respect to the critiques raised by Ferguson et al. (19–21), it is noteworthy that these researchers have conducted three rigorous longitudinal studies that have found no significant relationship between violent video game play and aggression. They attribute these noneffects in part to: (*i*) using measures of "serious" aggression (e.g., overt physical aggression), and (*ii*) including appropriate control covariates.

Ethnicity and Game Play

Some evidence exists supporting the potential of ethnicity and culture to moderate VGV effects. Anderson et al. (6) noted in their metaanalysis of aggressive behavior in longitudinal designs that the VGV effect was somewhat larger in Western than Eastern cultures and this difference approached statistical significance ($P = 0.07$). At the same time, in these comparisons cultural differences were confounded with variation in research designs, such that "it was unclear whether the difference should be attributed to cultural differences in vulnerability or to the use of different measures" (6).

The potential for ethnicity to moderate the effects of video game exposure on aggression was corroborated by Ferguson (15) in his own recent metaanalysis. In that work, Ferguson found a statistically significant association between exposure to video

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Digital Media and Developing Minds," held October 14–16, 2015, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA. The complete program and video recordings of most presentations are available on the NAS website at www.nasonline.org/Digital_Media_and_Developing_Minds.

Author contributions: A.T.P., J.D.S., and J.G.H. designed research; A.T.P., J.D.S., and J.G.H. performed research; A.T.P. and J.G.H. analyzed data; and A.T.P., J.D.S., and J.G.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: jay.g.hull@dartmouth.edu.

Research Report

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
XX(X) 1–4

DOI: 10.1177/0956797610368064
http://sagepub.com
SAGE

Mere Visual Perception of Other People's Disease Symptoms Facilitates a More Aggressive Immune Response

Mark Schaller, Gregory E. Miller, Will M. Gervais, Sarah Yager, and Edith Chen
University of British Columbia

Abstract

An experiment (N = 28) tested the hypothesis that the mere visual perception of disease-connoting cues promotes a more aggressive immune response. Participants were exposed either to photographs depicting symptoms of infectious disease or to photographs depicting guns. After incubation with a model bacterial stimulus, participants' white blood cells produced higher levels of the proinflammatory cytokine interleukin-6 (IL-6) in the infectious-disease condition, compared with the control (guns) condition. These results provide the first empirical evidence that visual perception of other people's symptoms may cause the immune system to respond more aggressively to infection. Adaptive origins and functional implications are discussed.

Keywords

disease, health, immunity, perception, threat

Received 9/11/09; Revision accepted 10/30/09

People are sensitive to visual stimuli connoting the potential presence of infectious pathogens in others. These stimuli include anomalous morphological and behavioral characteristics (e.g., skin discolorations, sneezing) that suggest infection with disease-causing microorganisms. When perceived, these stimuli trigger psychological responses—such as disgust and the activation of aversive cognitions into working memory—that inhibit interpersonal contact (e.g., Curtis, Aunger, & Rabie, 2004; Oaten, Stevenson, & Case, 2009; Park, Faulkner, & Schaller, 2003; Park, Schaller, & Crandall, 2007). These perceptual processes are part of an integrated set of psychological mechanisms that facilitate prophylactic behavioral defense against pathogens—a sort of *behavioral immune system* (Schaller & Duncan, 2007). Previously unexplored, however, is the intriguing possibility that these processes might also have an influence on the real immune system.

In a recent review article on disgust as a disease-avoidance mechanism, Oaten et al. (2009) suggested that "immune function, especially the innate (i.e., rapid) component, may be directly mobilized by cues that are disgust-evoking," but also noted that "as yet there are no data in humans to confirm or refute this possibility" (p. 315). Here, we report a study that empirically tested (and supports) the specific hypothesis that mere visual perception of other people's disease-connoting cues can cause the immune system to respond more vigorously to microbial stimuli that connote infection.

This hypothesis is plausible on functional grounds. Visual perception of other people's apparent symptoms of infection implies one's own immediate vulnerability to pathogen infection. To the extent that visual perception of such stimuli influences perceivers' own immune functioning (by causing perceivers' immune cells to respond more aggressively if, or when, such infection occurs), this response phenomenon may reduce the likelihood of the infection's becoming established.

The hypothesis is plausible on mechanistic grounds as well. There is abundant evidence that immune responses (e.g., the production of proinflammatory cytokines) can be facilitated by stressful psychological experiences. These effects are mediated by hormones such as cortisol and norepinephrine, which are released when people appraise situations as threatening, and subsequently bind to receptors on immune cells (Cohen, Doyle, & Skoner, 1999; Kiecolt-Glaser et al., 2003;

Corresponding Author:
Mark Schaller, Department of Psychology, University of British Columbia, 2136 West Mall, Vancouver, British Columbia V6T 1Z4, Canada
E-mail: schaller@psych.ubc.ca

Table 1. Mean Stimulated Production of Interleukin-6 (IL-6) and Self-Reported Mood Before and After the Guns and Disease Slide Shows

Measure	Guns slide show	Disease slide show
Stimulated IL-6		
Pretest (pg/ml)	32,002 (29,974)	22,320 (14,672)
Posttest (pg/ml)	33,964 (30,725)	26,814 (15,771)
Change (pg/ml)	1,962 (3,790)	4,494 (8,249)
Change (%)	6.62 (20.51)	23.62 (31.74)
Self-reported mood		
Stressed	1.57 (0.94)	1.24 (0.96)
Relaxed	1.62 (1.18)	1.67 (1.13)
Scared	1.38 (1.11)	0.88 (0.89)
Disgusted	1.52 (1.19)	1.64 (1.17)

Note: Standard deviations are given in parentheses. Mood was assessed after the slide show only.

disease condition (see Table 1). Does this difference reflect a failure of randomization? It appears not. In addition to the primary measures described earlier, all participants completed a battery of questionnaires assessing dispositional tendencies, including the Big Five personality traits (agreeableness, conscientiousness, extraversion, neuroticism, and openness), as well as six specific traits relevant to perceptions of threat and disease (e.g., perceived vulnerability to disease, health locus of control). On none of these traits was there a significant difference between subjects in the guns and disease conditions (all $ps \geq .10$). (Nor did any of these traits significantly predict changes in stimulated IL-6; because of these noneffects, the trait measures are not discussed further in this article.) Furthermore, the difference between slide-show conditions in pretest levels of stimulated IL-6 was nonsignificant ($p = .288$), and pretest values of stimulated IL-6 had no meaningful relation to the percentage of change in stimulated IL-6 ($rs = –.03$ and $–.18$ in the guns and disease conditions, respectively; both $ps > .54$). Most important, the significant between-conditions difference in relative pretest-to-posttest change in stimulated IL-6 (revealed by the 2 × 2 ANOVA reported earlier) remained significant even when we statistically controlled for pretest values of stimulated IL-6 ($p = .004$).

Can this latter difference be attributed to greater subjective

Research Report

Guns, Testosterone, and Aggression

An Experimental Test of a Mediational Hypothesis

Jennifer Klinesmith, Tim Kasser, and Francis T. McAndrew

Knox College

ABSTRACT—*We tested whether interacting with a gun increased testosterone levels and later aggressive behavior. Thirty male college students provided a saliva sample (for testosterone assay), interacted with either a gun or a children's toy for 15 min, and then provided another saliva sample. Next, subjects added as much hot sauce as they wanted to a cup of water they believed another subject would have to drink. Males who interacted with the gun showed significantly greater increases in testosterone and added more hot sauce to the water than did those who interacted with the children's toy. Moreover, increases in testosterone partially mediated the effects of interacting with the gun on this aggressive behavior.*

Substantial evidence suggests that aggression can be increased by the presence of weapons in the environment and by the hormone testosterone. Several studies show that the presence of aggressive environmental cues such as weapons can increase the accessibility of hostile, aggressive thoughts and lead to more aggressive behavior (Anderson, Benjamin, & Bartholow, 1998; Bartholow, Anderson, Carnagey, & Benjamin, 2005; Berkowitz & LePage, 1967; Bettencourt & Kernahan, 1997; Killias & Haas, 2002). Regarding testosterone, in animal species ranging from chickens to monkeys, the injection of this hormone increases aggressiveness and social dominance behavior, regardless of whether the animals are males or females (Ellis, 1986); in humans, however, the results are more mixed, with many laboratory and field studies revealing strong positive relations between testosterone and levels of restlessness, tenseness, and tendency toward violence (Archer, 1994; Campbell, Muncer, & Odber, 1997; Dabbs, Carr, Frady, & Riad, 1995; Dabbs, Jurkovic, & Frady, 1991) and other studies failing to replicate such effects (Archer, 1991; Archer, Birring, & Wu, 1998; O'Connor, Archer, Hair, & Wu, 2001; Rowe, Maughan, Worthman, Costello, & Angold, 2004).

Surprisingly, we were unable to find any studies that examined whether testosterone and the presence of a weapon might work together to increase aggressive behavior. Perhaps the presence of a stimulus such as a gun triggers increases in testosterone levels, which in turn increase aggressive behavior. Such a chain of events would be predicted by the *challenge hypothesis* developed by Wingfield, Hegner, Dufty, and Ball (1990) to explain aggressive behavior in male pair-bonded birds. According to this hypothesis, testosterone rises in response to situational cues that represent either a threat to a male's status or a signal that competition with other males is imminent; such increases in testosterone then facilitate whatever competitive behaviors (including potentially aggressive responses) are necessary for meeting the challenge. The challenge hypothesis has been supported by studies across a wide range of vertebrate species (Cavigelli & Pereira, 2000; Ferree, Wikelski, & Anderson, 2004; Hirschenhauser, Taborsky, Oliveira, Canario, & Oliveira, 2004; Muller & Wrangham, 2004); most studies in humans have focused on how males' testosterone levels rise and fall depending on success or failure in competitions (Archer, 1991; Booth, Shelley, Mazur, Tharp, & Kittok, 1989; Gladue, Boechler, & McCaul, 1989; Mazur, Booth, & Dabbs, 1992; Mazur & Lamb, 1980) or in response to insults (Cohen, Nisbett, Bowdle, & Schwarz, 1996; see Archer, 2006, for a review of the applicability of the challenge hypothesis to humans).

In this study, we examined whether the presence of a gun (vs. a control object) might act as a stimulus signaling competition and a threat to status; if so, according to the challenge hypothesis, it should cause increases in males' testosterone levels, which in turn should increase their aggressive behavior. We assessed males' testosterone levels both before and after interacting with a gun or a children's toy; to measure aggression, we adapted a method developed by Lieberman, Solomon, Greenberg, and

Address correspondence to Tim Kasser, Box K-83, Knox College, Galesburg, IL 61401, e-mail: tkasser@knox.edu.

McGregor (1999) that gives subjects the opportunity to anonymously put hot sauce in a cup of water that they believe another person will have to drink. We hypothesized that males who interacted with the gun would show both a greater increase in testosterone levels and more aggression than would males who interacted with the children's toy. We also hypothesized that changes in testosterone levels would be correlated with aggression levels and would indeed mediate the effects of interacting with a gun on later aggressive behavior.

METHOD

Subjects

Subjects were 30 male college students (age range: 18–22) who received extra course credit or a small monetary reward for their participation. All subjects were run during the afternoon or early evening.

Procedure and Materials

When recruited, subjects were informed that the study would examine taste sensitivity in males and that they would therefore need to provide saliva for hormone analysis; subjects were asked not to eat, drink, smoke, or brush their teeth for 1 hr prior to testing in order to minimize impurities in the saliva samples. When subjects arrived at the lab, a female experimenter confirmed that the subjects had indeed followed these instructions before she administered consent procedures. Next, participants provided an approximately 6-ml sample of saliva by spitting into a cup; this saliva was used to assess baseline, or Time 1, testosterone levels.

All subjects were then led into a room containing a television, a chair, and a table with an object and some paper on it. For experimental subjects, the object was a pellet gun identical in size, shape, and feel to a Desert Eagle automatic handgun; for control subjects, the object was the children's game Mouse Trap™. Subjects were told that the study was investigating whether taste sensitivity was associated with the attention to detail required for creating instructions concerning the object. Subjects were therefore asked to spend 15 min handling the object and writing a set of instructions about how to assemble and disassemble it; a drawing of the object was also provided for subjects to label the object's parts. The handgun and children's game were similar in number and complexity of parts.

After 15 min, the experimenter reentered the room, asked the subject to stop working on the instructions, and obtained a Time 2 saliva sample from the subject. The subject was told he would next perform the taste-sensitivity portion of the study. He was given a cup filled with 85 g of water and a single drop of Frank's Red Hot Sauce. The subject was told that the sample had been prepared by a previous subject, was instructed to take a sip of the sample, and was then asked to rate the taste of the sample on a scale provided.

The experimenter left and then returned with a tray containing a cup of 85 g of water, a nearly full bottle of Frank's Red Hot Sauce, and a lid. The subject was asked to prepare a sample for the next subject by placing as much hot sauce in the water as he wanted. He was assured that neither the person who drank it nor the experimenter would know how much hot sauce he had put in the water, as the lid was to be put on the cup after the hot sauce was added. The experimenter then left the room, and the subject signaled when he was finished adding the hot sauce. (Throughout this process, the gun or the game remained in the room.) The cup was then removed from the room, and the experimenter weighed it again to obtain a measure of the amount of hot sauce, in grams, the subject had added to the water. This served as our primary measure of aggression (see Lieberman et al., 1999).

Because of the potentially arousing nature of the experiment, we wanted to ensure that all subjects were reasonably calm when they left the lab. Therefore, all subjects next watched a relaxing video of nature scenes and classical music. Given that subjects had been deceived, we next debriefed them, emphasizing that they should not feel badly about any aggressive behavior they exhibited. Interestingly, several subjects were disappointed when told that the sample of hot sauce and water they had prepared would not actually be given to the next subject. No subjects expressed suspicion as to the true nature of the study.

Testosterone Levels

Time 1 and Time 2 saliva samples were stored for 24 hr at room temperature, centrifuged, and then frozen at −20 °C until the time of the assay (Erikkson & Von Der Pahlen, 2002). The samples were then brought to room temperature, transferred to Eppendorf tubes, centrifuged for 15 min at 3,000 rpm to remove debris, and then assayed in duplicate using a commercially available microwell kit for testosterone level (Salimetrics, LLC, State College, PA). All samples were assayed in house in a single batch using a standard radioimmunoassay (RIA) procedure under the supervision of an experienced RIA technician; at both Time 1 and Time 2, the duplicates were averaged to yield our measures of testosterone level. The intra-assay coefficient of variation for subjects was 5.3%, and the sensitivity of the assay was less than 1.5 pg/ml from zero for men. Mean Time 1 and Time 2 testosterone levels were 222.59 pg/ml (SD = 97.17) and 253.92 pg/ml (SD = 98.32), respectively. We subtracted each subject's Time 1 level from his Time 2 level to obtain a measure of change in testosterone.

RESULTS

Our first hypothesis was confirmed: Subjects who interacted with the handgun showed a greater increase in testosterone from Time 1 to Time 2 (mean change = 62.05 pg/ml, SD = 48.86) than did those who interacted with the children's game (mean

Pre-post, 2 groups Repeated Measures Anova

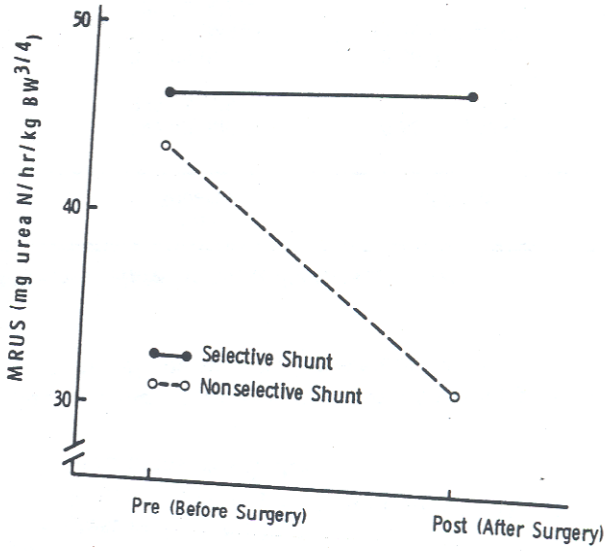
Brogan-Kutner example

Stat week [redacted]

© The American Statistician, November 1980, Vol. 34, No. 4

1. Pre and Post Maximal Rate of Urea Synthesis Level (mg urea N/hr/kg $BW^{3/4}$) and Sample Cell Means, by Group

Group	Subject	Pre	Post
Selective Shunt (new operation)	1	51	48
	2	35	55
	3	66	60
	4	40	35
	5	39	36
	6	46	43
	7	52	46
	8	42	54
Mean		$\hat{\mu}_{11} = 46.375$	$\hat{\mu}_{12} = 47.125$
Nonselective Shunt (standard operation)	9	34	16
	10	40	36
	11	34	16
	12	36	18
	13	38	32
	14	32	14
	15	44	20
	16	50	43
	17	60	45
	18	63	67
	19	50	36
	20	42	34
	21	43	32
Mean		$\hat{\mu}_{21} = 43.538$	$\hat{\mu}_{22} = 31.462$



analysis on back

Growth Curves (Group) T=4

Bock, DR MSMBR text

EXAMPLE 7.1-1 (*Mixed-model analysis of vocabulary growth*) Data for this example are drawn from test results on file in the Records Office of the Laboratory School of the University of Chicago. They consist of scores, obtained from a cohort of pupils at the eighth through eleventh grade level, on alternative forms of the vocabulary section of the Cooperative Reading Tests [Davis, 1950]. Since these data cover an age range in which physical growth is beginning to decelerate, it is of interest to inquire whether a similar deceleration can be observed in the acquisition of new vocabulary.

Table 7.2-5 MIXED-MODEL ANALYSIS OF VARIANCE OF SEX EFFECTS IN THE VOCABULARY-SCALED SCORES

Source	df	ss	F	p
Constant	1	ssm = 1,644.90		
Sex	1	ssb = .85	.06	> .5
Occasions	3	ssc = 194.18		
Sex × occasions	3	ssbc = 2.79	1.12	> .1
Subjects within groups	62	ssa = 873.00		
Occasions × subjects within groups	186	sse = 152.17		
Total	256	sst = 2,867.90		

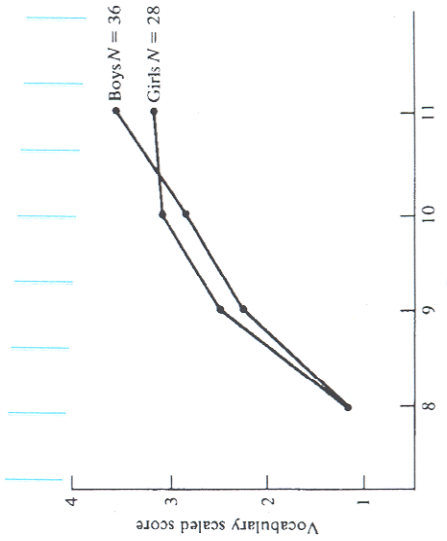


FIGURE 7.2-1
Average vocabulary scores of boys and girls in a cohort from the University of Chicago Laboratory School (longitudinal data).

In this section, *The American Statistician* publishes articles and notes of interest to teachers of the first mathematical statistics course and of applied statistics courses. To be suitable for this section, articles and notes should be useful to a substantial number of teachers of such a course or should have the potential for fundamentally affecting the way in which the course is taught.

Comparative Analyses of Pretest-Posttest Research Designs

DONNA R. BROGAN AND MICHAEL H. KUTNER*

Two common methods of analyzing data from a two-group pretest-posttest research design are (a) two-sample t test on the difference score between pretest and posttest and (b) repeated-measures/split-plot analysis of variance. The repeated-measures/split-plot analysis subsumes the t test analysis, although the former requires more assumptions to be satisfied. A numerical example is given to illustrate some of the equivalences of the two methods of analysis. The investigator should choose the method of analysis based on the research objective(s).

KEY WORDS: Repeated-measures/split-plot analysis; t test; Pretest-posttest designs.

1. INTRODUCTION

A common research design is the two-group pretest/posttest design with one dependent variable where subjects are not matched and may or may not be randomly assigned to the two groups (Cook and Campbell 1979). When the two groups are not formed by random assignment of subjects, a random sample from each of the two groups is necessary. This design can be extended to more than two groups; an example is the comparison of several different treatments with each other or with a control group in which each group is measured on a pretest and posttest.

The statistical analysis for these designs can be approached from several viewpoints. If the dependent variable is measured on an interval or ratio scale, a common analysis is to define a difference score for each subject (posttest minus pretest or vice versa) or a relative difference measure (the difference divided by the pretest) and then test the null hypothesis that the means or medians of the (relative) differences are equal for each group. In many cases the t test or analysis of variance is used, although nonparametric tests could also be used, for example, the Mann-Whitney U test, or the median test, or their analogs for more than two groups.

Covariance analysis, where the pretest score is used as the covariate, is another method used for analyzing this design. The difference score method is essentially a special case of the analysis of covariance where the regression coefficient of the posttest on the pretest is assumed to equal unity. Neter and Wasserman (1974, p. 717) and Cox (1958, pp. 55–56) point out that if the common slope is not near one the covariance analysis probably will be better than the difference score analysis. We note that when an experimental group is to be compared to a control group, it is often likely that inequality of slopes will prevail among groups, thus violating an assumption of the analysis of covariance. Bock (1975, Sec. 7.3) compares the interpretation of a difference-score analysis and covariance analysis and suggests guidelines regarding which analysis to use.

Still another method of analyzing this design is to view the pretest and posttest as a repeated-measures/split-plot design or as a profile of two measurements for each subject. Repeated-measures/split-plot designs are discussed in detail by Winer (1971) and Steel and Torrie (1980), whereas both repeated measures and profile analysis are discussed in Morrison (1976, Secs. 4.5, 4.6, and 5.6).

This article illustrates some of the equivalences and differences between the difference score analysis and the repeated-measures/split-plot or profile analysis. The numerical example and major discussion are for a two-group pretest/posttest design where subjects are not matched. Concluding remarks indicate how the results can be extended easily to more than two groups.

2. A NUMERICAL EXAMPLE

We consider data from Rikkers et al. (1978), who report results of a prospective randomized surgical trial allocating cirrhotic patients who had bled from varices to either a nonselective shunt (standard operation) or to a selective shunt (new operation). The dependent variable is the maximal rate of urea synthesis (MRUS), which is a quantitative test of liver function. Poor liver function is associated with a low MRUS value. MRUS was measured preoperatively and early postoperatively in eight selective shunt patients and thirteen nonselective shunt patients. The purposes of the study were to assess preoperatively the comparability of the selective and the nonselective groups and to longitudinally evaluate the change in liver function

* Donna R. Brogan is Professor and Michael H. Kutner is Associate Professor in the Statistics and Biometry Department, School of Medicine, Emory University, Atlanta, GA 30322. Work on this article was partially supported by NCI Contract No. CB–74101 and USPHS Grant No. RR39.

1. Pre and Post Maximal Rate of Urea Synthesis Level (mg urea N/hr/kg $BW^{3/4}$) and Sample Cell Means, by Group

Group	Subject	Pre	Post
Selective Shunt (new operation)	1	51	48
	2	35	55
	3	66	60
	4	40	35
	5	39	36
	6	46	43
	7	52	46
	8	42	54
Mean		$\hat{\mu}_{11} = 46.375$	$\hat{\mu}_{12} = 47.125$
Nonselective Shunt (standard operation)	9	34	16
	10	40	36
	11	34	16
	12	36	18
	13	38	32
	14	32	14
	15	44	20
	16	50	43
	17	60	45
	18	63	67
	19	50	36
	20	42	34
	21	43	32
Mean		$\hat{\mu}_{21} = 43.538$	$\hat{\mu}_{22} = 31.462$

of the two groups. Table 1 reports the MRUS values for each patient for the preoperative and postoperative periods and the respective cell means.

For completeness Table 2 displays the standard repeated-measures analysis of variance table (analysis of means method); in this example the total number of subjects, n, is 21. The hypotheses of interest to the researchers were the interaction test and the simple effects test on equality of preoperative population means between groups. The test for interaction is significant ($F = 11.36$ with 1 and 19 df, $p < .005$); therefore, it is concluded that the pre/post average change in the nonselective group is significantly different from the pre/post average change in the selective group (see the figure). In the presence of a significant interaction effect it is generally of interest to test simple effects rather than main effects (Winer 1971, p. 529). The Bonferroni multiple-comparison procedure (see Neter and Wasserman 1974) was adopted to test the following contrasts:

$$(\mu_{12} - \mu_{11}) - (\mu_{22} - \mu_{21}) = 0$$

$$\mu_{12} - \mu_{11} = 0$$

$$\mu_{22} - \mu_{21} = 0$$

and

$$\mu_{11} - \mu_{21} = 0.$$

Using an experiment wise error rate of .05, we conclude that the interaction effect is significant and μ_{21} is significantly greater than μ_{22}. Therefore, significant deterioration of liver function occurred in the nonselective patients between preoperative and early postoperative evaluation periods, whereas the selective group had no apparent deleterious effect. Two points are worth noting in the example just cited: (a) The equality of slopes test using the preoperative MRUS values as a covariate is rejected ($p < .02$); and (b) the significant interaction effect requires special handling when testing the last contrast since, for the pretest level, we have a two-group experiment in which there are no repeated measures. Therefore, the appropriate error term for this type of comparison is MS (within cell). For a more extensive coverage of this point the reader is referred to Winer (1971, pp. 529–532).

3. REPEATED-MEASURES ANALYSIS

We now discuss the statistical properties of the repeated-measures analysis of variance for this example and compare it with the statistical properties of the difference score analysis. Using the model proposed by Winer (1971, p. 519), we have

$$X_{ijk} = \mu + \alpha_i + \Pi_{k(i)} + \beta_j + \alpha\beta_{ij} + \beta\Pi_{jk(i)} + \epsilon_{m(ijk)} \quad (3.1)$$

$j = 1, 2$ (pretest = 1, posttest = 2),

$i = 1, 2$ (group 1 = 1, group 2 = 2),

$k = 1, 2, \ldots, n_i, \quad m = 1,$

where X_{ijk} is the observed value of subject k within group i at time j,

μ is the overall mean,
α_i is the effect of group i,
$\Pi_{k(i)}$ is the effect of subject k nested within group i,
β_j is the effect of the repeated-measures variable j (*i.e.*, pretest and posttest),
$\alpha\beta_{ij}$ is the interaction of group i with level j of the repeated measures factor,

2. Repeated-Measures Analysis of Variance for Maximal Rate of Urea Synthesis Level

Source of Variation	df	Sum of Squares	Mean Squares	F Ratio
Between Subjects	20 (n − 1)			
Groups	1	847.48	847.48 (MS_G)	3.63 (MS_G/MS_E)
Subjects Within Groups	19 (n − 2)	4440.00	233.68 (MS_E)	
Within Groups	21 (n)			
Pre/Post	1	317.69	317.69 (MS_P)	8.86 (MS_P/MS_{PE})
Groups x Pre/Post	1	407.41	407.4 (MS_{GP})	11.36 (MS_{GP}/MS_{PE})
(Pre/Post) x Subjects Within Groups	19 (n − 2)	681.21	35.85 (MS_{PE})	

and

$\beta\Pi_{jk(i)}$ is the interaction of subject k within group i with level j of the repeated-measures factor.

The following constraints are imposed on the parameters:

$$\alpha_{\cdot} = \beta_{\cdot} = \alpha\beta_{i\cdot} = \alpha\beta_{\cdot j} = 0,$$

where

$$\alpha\beta_{\cdot j} = \sum_i \alpha\beta_{ij}, \text{ and so on.} \qquad (3.2)$$

In the design under discussion, the repeated-measures factor and the group factor are each at two levels.

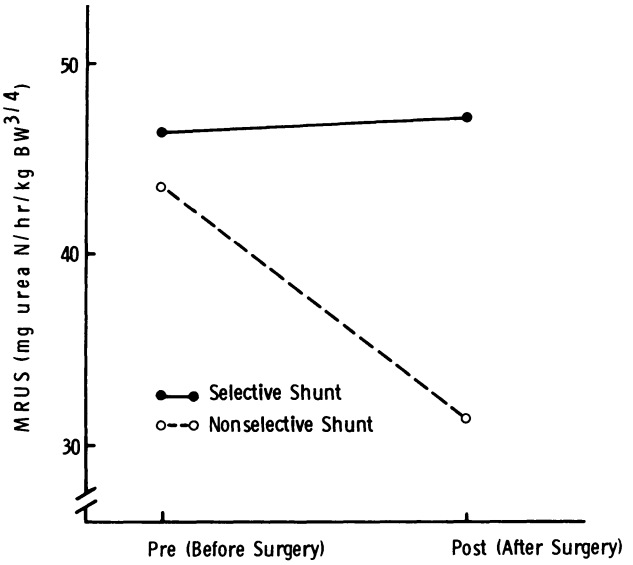
The general analysis of variance also is indicated in Table 2, where n is the total number of subjects. Note that it is not necessary for each group to contain the same number of subjects. Assuming the group factor and the pre/post factor to be fixed effects, Winer (1971) shows that the appropriate F tests are as indicated in the F ratio column of Table 2.

It is worth noting exactly what null hypotheses are tested in Table 2. The ratio MS_G/MS_E tests the null hypothesis that there is no *group main effect*. This is equivalent to testing whether the *sum* of the pretest and posttest observations on each subject has the same population mean in the two groups. The ratio MS_P/MS_{PE} tests the null hypothesis that there is no *pre/post main effect* and is equivalent to testing whether the population mean of the pretest observations is the same as the population mean of the posttest observations. The ratio MS_{GP}/MS_{PE} tests the null hypothesis that there is no *interaction* between the group main effect and the pre/post main effect. This ratio also tests whether the difference between pretest and posttest observations has the same population mean in both groups. This is the test many researchers are interested in when using this research design, since they often wish to assess whether a treatment has had any effect upon an experimental group. Note that this F test has $(1, n - 2)$ df, which will correspond to the t test with $(n - 2)$ df, as discussed in the next section.

Two assumptions are required to arrive at the F tests indicated in Table 2 (Winer 1971).

1. The pretest and posttest population variance-covariance matrices for each group are assumed equal.
2. The random effects $\Pi_{k(i)}$, $\beta\Pi_{jk(i)}$, and $\epsilon_{m(ijk)}$ from the model in (3.1) are all independently and normally distributed with mean zero and variances σ_Π^2, $\sigma_{\beta\Pi}^2$, and σ_E^2, respectively.

Assumption (1), equality of the variance-covariance matrices, implies two other results worth noting. First, the variation of subjects within the two groups is homogeneous. That is, if each subject's pretest and posttest observations are added together, this sum has the same population variance in both groups. This allows pooling over groups to calculate SS_E. Second, the variation of the interaction of subject and the pre/post factor is homogeneous for the two groups. That is, if the difference score between pretest and posttest is defined for each subject, the population variance of the difference scores is the same for both groups. This allows pooling over groups to calculate SS_{PE}.



Mean Pre and Post Maximal Rate of Urea Synthesis Level (MRUS) by Type of Surgery

4. DIFFERENCE-SCORE ANALYSIS

Using model (3.1) and forming a difference score d_{ik} for each subject k nested in group i yields

$$d_{ik} = X_{i1k\cdot} - X_{i2k}$$
$$= (\beta_1 - \beta_2) + (\alpha\beta_{i1} - \alpha\beta_{i2}) + (\beta\Pi_{1k(i)} - \beta\Pi_{2k(i)}) + (\epsilon_{m(i1k)} - \epsilon_{m(i2k)}). \qquad (4.1)$$

The term $(\alpha\beta_{i1} - \alpha\beta_{i2})$ is a parameter associated with group i and measures the "effect" of group i on the difference score d_{ik}. The null hypothesis to be tested is H_o: $\alpha\beta_{11} - \alpha\beta_{12} = \alpha\beta_{21} - \alpha\beta_{22}$ or H_o: $\alpha\beta_{11} - \alpha\beta_{12} = \alpha\beta_{21} - \alpha\beta_{22} = 0$. The difference scores in (4.1) can be viewed as a one-way classification model in which the error term is the sum of the following two terms:

$$(\beta\Pi_{1k(i)} - \beta\Pi_{2k(i)}) \quad \text{and} \quad (\epsilon_{m(i1k)} - \epsilon_{m(i2k)}).$$

If we assume this error term and also homogeneous variances for the two groups, an appropriate test statistic is the Student t test for two independent samples with $(n - 2)$ df.

5. COMPARISON OF THE TWO ANALYSES

The following three results are useful computationally and can be verified easily with the example from Table 1.

1. If the sum of the pretest and posttest is formed for each subject and a two-sample t test is used to compare the group means of the sum, then the calculated $t = 1.904$ and is the square root of the F test for Groups in Table 2 with (1,19) df.

2. If the difference between the pretest and the posttest is formed for each subject and a two-sample t test is used to compare the group means of these differences, then the calculated $t = 3.371$ and is the square root of the F test for Groups x (pre/post) in Table 2 with (1,19) df.
3. If all 21 subjects are considered to be in one group, then the t statistic to test the null hypothesis that the mean difference score is zero has 20 df and equals 3.158. From Table 2, if we reanalyze the within-subjects component by assuming that the Group x (pre/post) interaction is zero and SS_{GP} is pooled with SS_{PE}, then the F test for the main effect pre/post yields $F = 9.97$ with (1,20) df, which equals the square of the preceding t statistic.

These results demonstrate that the various F tests in the repeated-measures analysis of variance can be obtained by using simple t tests on linear combinations of the pre and post scores. It can be shown algebraically that interpretations (1), (2), and (3) of the F tests hold in the particular research design discussed in this article, that is, a two-group pretest/posttest design. In fact, the numerical operations in (1) and (2) of summing and differencing the pretest/posttest observations are used by the latest version of the BMD P2V program in calculating sums of squares in repeated-measures designs (Dixon and Brown 1979).

Since the difference-score analysis is embedded in the repeated-measures analysis, the repeated-measures analysis provides more information about the data at hand. Fewer assumptions, however, are required in the difference-score analysis. The difference-score analysis assumes only homogeneous variances for the difference scores and a normally distributed error term with mean zero. It is easy to show that if the assumptions of the repeated-measures analysis are satisfied, then the assumptions of the difference-score analysis are also met. However, the converse is not true.

In our experience, the researcher rarely is interested in only the interaction test, that is, the difference-score analysis. Furthermore, simple effects are commonly of interest even in the no interaction effect experiments. Therefore, we advocate the use of the repeated-measures/split-plot analysis in most instances. However, we urge the user to empirically validate the underlying assumptions.

In the example discussed in Section 2, the researchers would have been interested in assessing the significance of the main effect time (pre vs. post) *if* the Groups x pre/post interaction had been nonsignificant. That is, the nonsignificant interaction would have indicated that the two groups did not differ significantly on their MRUS difference scores. The pre/post main effect test would then indicate whether the MRUS difference score in both groups was significantly different from zero, that is, whether the treatment did or did not effect both groups.

6. GENERALIZATION OF FINDINGS

If there are more than two groups, similar results can be obtained. The difference-score analysis would no longer be performed by a t test but by an F test in a one-way analysis of variance. A further extension can be made where the different groups being compared may be defined by several factors in a factorial design.

However, there is no logical extension of this discussion to more than two levels of the repeated-measures factor since a simple difference score analysis would no longer be appropriate.

[*Received July 1977. Revised June 1980.*]

REFERENCES

Bock, R. Darrell (1975), *Multivariate Statistical Methods in Behavioral Research*, New York: McGraw-Hill Book Co.

Cook, Thomas D., and Campbell, Donald T. (1979), *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Chicago: Rand McNally.

Cox, D.R. (1958), *Planning of Experiments*, New York: John Wiley & Sons.

Dixon, Wilfred J., and Brown, Morton B. (eds.) (1979), *Biomedical Computer Programs P-Series*, Berkeley: University of California Press.

Morrison, Donald F. (1976), *Multivariate Statistical Methods* (2nd ed.), New York: McGraw-Hill Book Co.

Neter, John, and Wasserman, William (1974), *Applied Linear Statistical Models*, Homewood, Ill.: Richard D. Irwin.

Rikkers, Layton F., Rudman, Daniel, Galambos, John T., Fulenwider, J. Timothy, Milliken, William J., Kutner, Michael H., Smith, Robert B., Salam, Atef A., Sones, Peter J., and Warren, W. Dean (1978), "A Randomized, Controlled Trial of the Distal Spenorenal Shunt," *Annals of Surgery*, 188, 271–282.

Steel, Robert G.D., and Torrie, James H. (1980), *Principles and Procedures of Statistics* (2nd ed.), New York: McGraw-Hill Book Co.

Winer, B.J. (1971), *Statistical Principles in Experimental Design* (2nd ed.), New York: McGraw-Hill Book Co.

LONG FORM

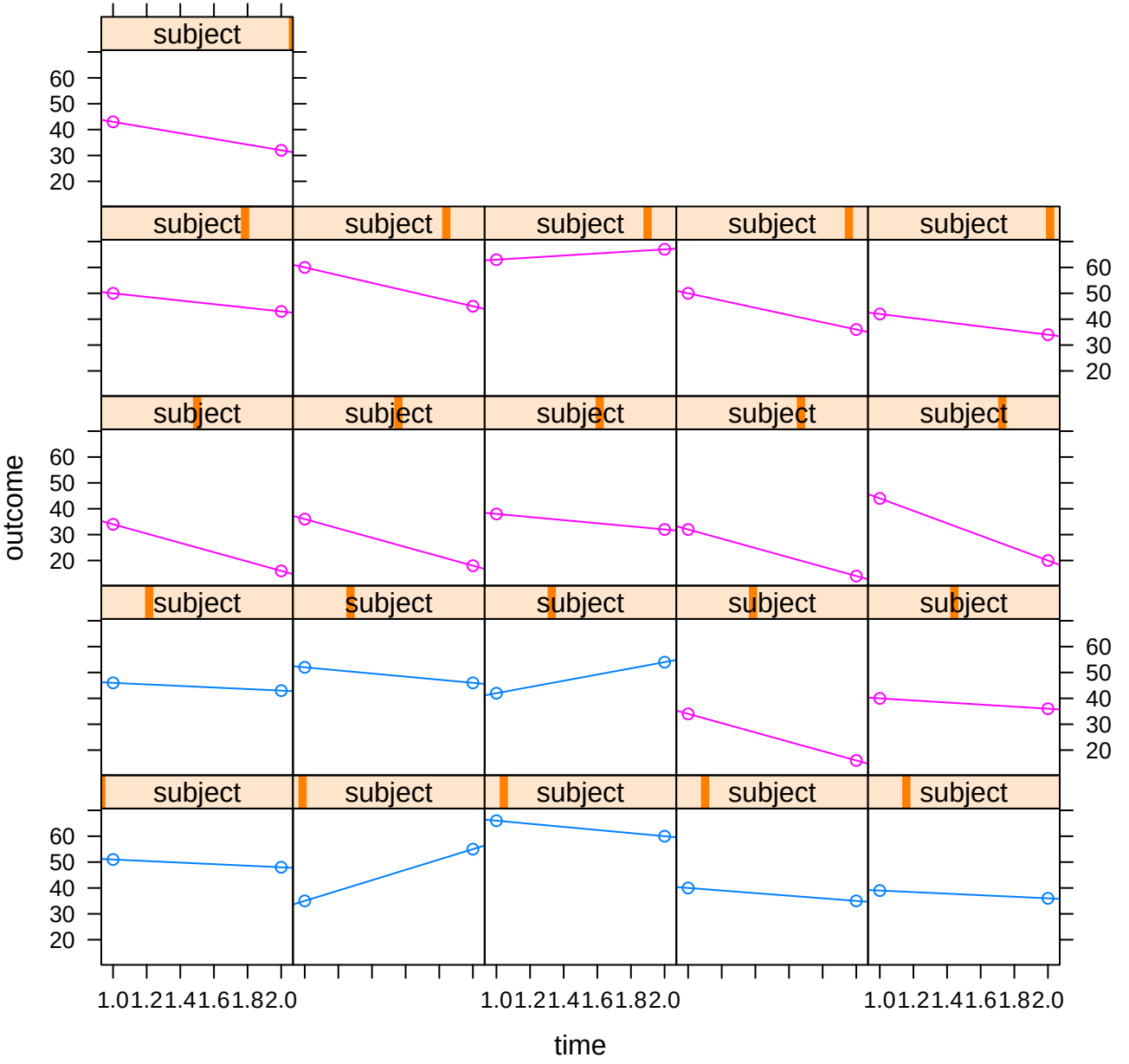
```
> bk
   method prepost urea subj
1       1       1   51    1
2       1       2   48    1
3       1       1   35    2
4       1       2   55    2
5       1       1   66    3
6       1       2   60    3
7       1       1   40    4
8       1       2   35    4
9       1       1   39    5
10      1       2   36    5
11      1       1   46    6
12      1       2   43    6
13      1       1   52    7
14      1       2   46    7
15      1       1   42    8
16      1       2   54    8
17      2       1   34    9
18      2       2   16    9
19      2       1   40   10
20      2       2   36   10
21      2       1   34   11
22      2       2   16   11
23      2       1   36   12
24      2       2   18   12
25      2       1   38   13
26      2       2   32   13
27      2       1   32   14
28      2       2   14   14
29      2       1   44   15
30      2       2   20   15
31      2       1   50   16
32      2       2   43   16
33      2       1   60   17
34      2       2   45   17
35      2       1   63   18
36      2       2   67   18
37      2       1   50   19
38      2       2   36   19
39      2       1   42   20
40      2       2   34   20
41      2       1   43   21
42      2       2   32   21
```

```
> tapply(urea, list(method, prepost), mean)
          1        2
1 46.37500 47.12500
2 43.53846 31.46154
```

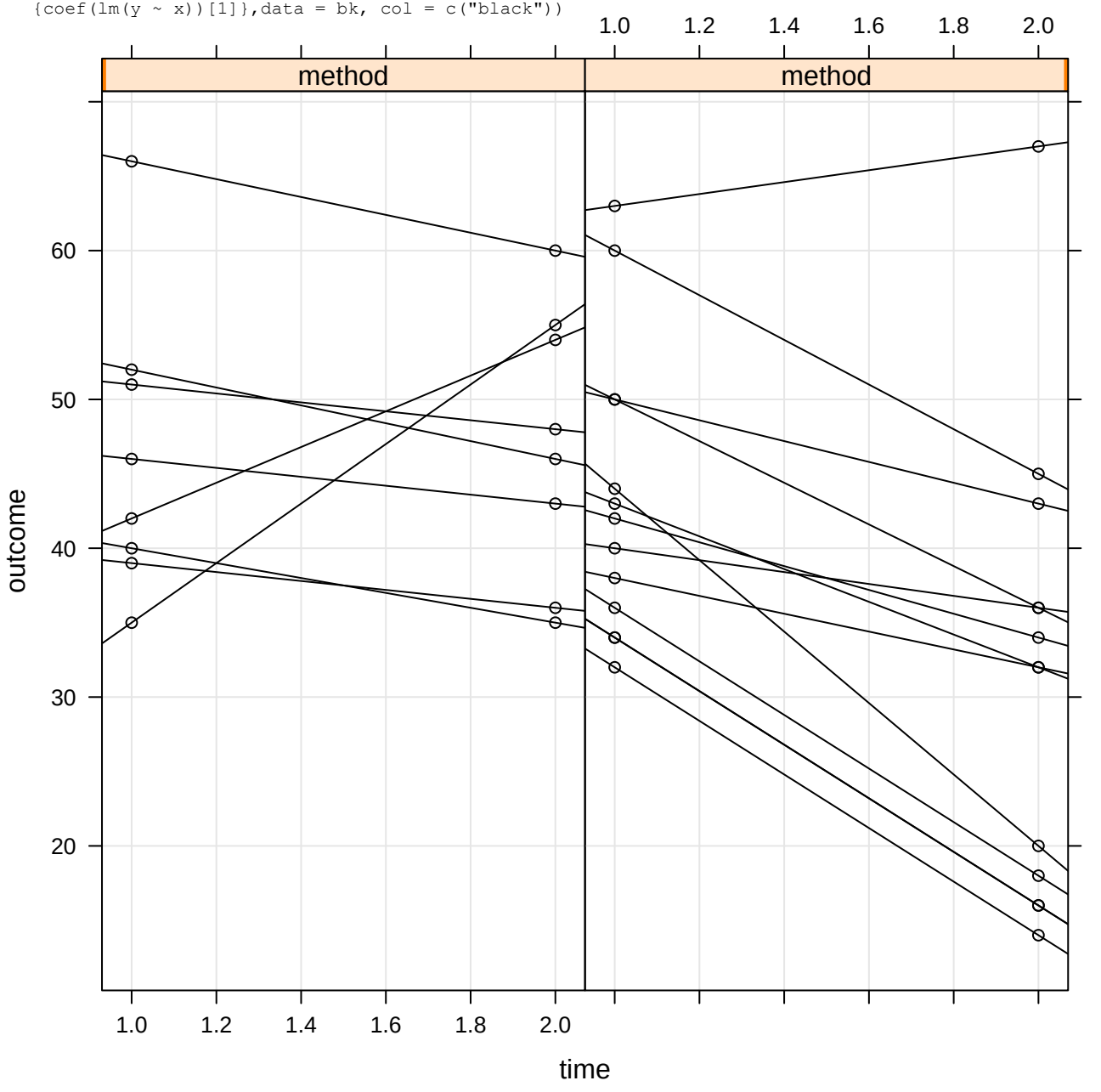
```
> bkrepaovW1 = aov(urea[method == "1"] ~ as.factor(prepo
> summary(bkrepaovW1)

as.factor(prepost[method == "1"])
as.factor(subj[method == "1"])
as.factor(prepost[method == "1"]):as.factor(subj[method
> bkrepaovW2 = aov(urea[method == "2"] ~ as.factor(prepo
> summary(bkrepaovW2)

as.factor(prepost[method == "2"])
as.factor(subj[method == "2"])
as.factor(prepost[method == "2"]):as.factor(subj[method
> summary(bkrepaovBase)
                                       Df Sum Sq Mean Sq F
as.factor(prepost)                      1  542.9   542.9
as.factor(method)                       1  847.5   847.5
as.factor(prepost):as.factor(method)    1  407.4   407.4
Residuals                              38 5121.2   134.8
```

LONG FORM

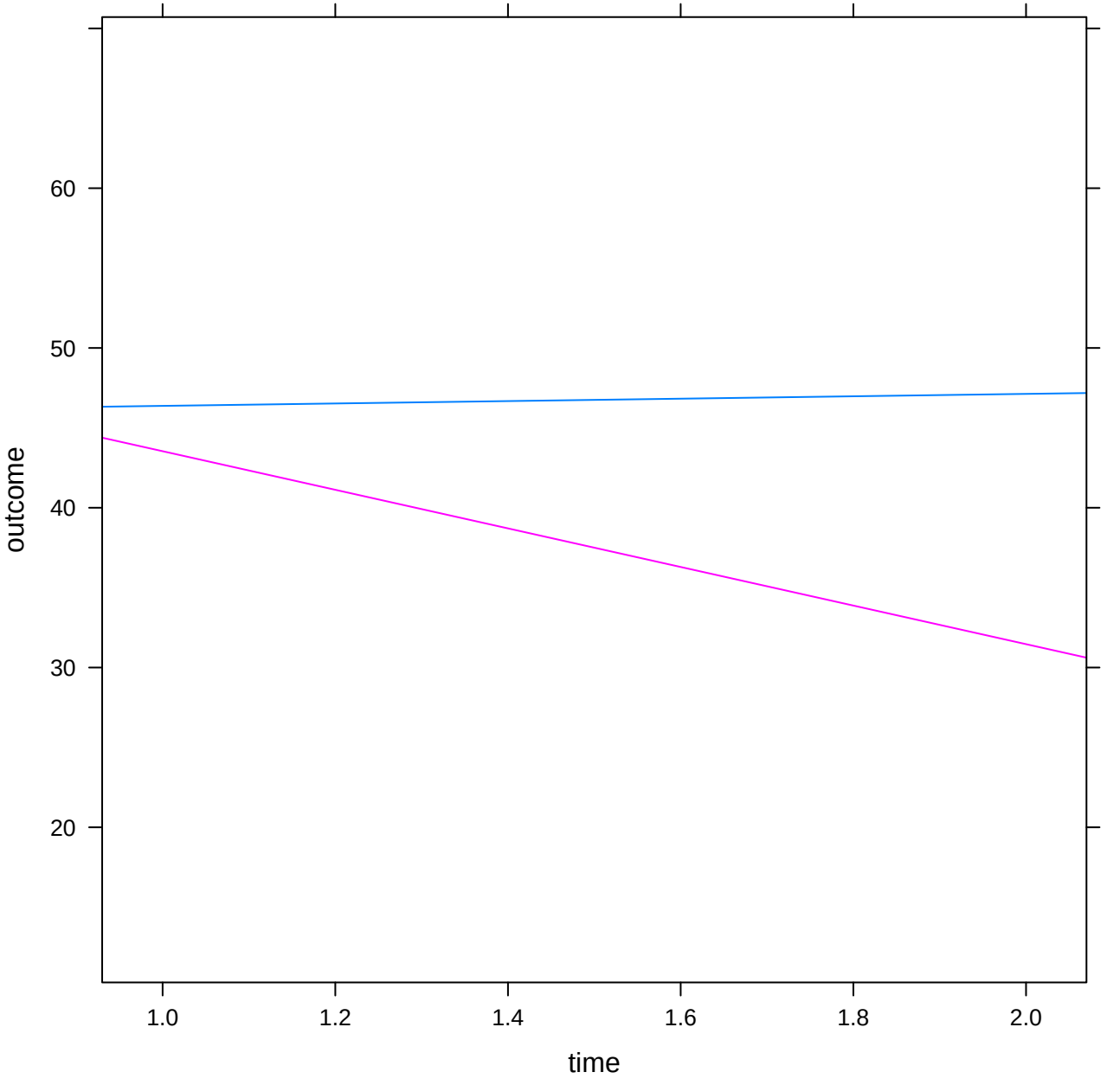
```
> bk
   method prepost urea subj
1       1       1   51    1
2       1       2   48    1
3       1       1   35    2
4       1       2   55    2
5       1       1   66    3
6       1       2   60    3
7       1       1   40    4
8       1       2   35    4
9       1       1   39    5
10      1       2   36    5
11      1       1   46    6
12      1       2   43    6
13      1       1   52    7
14      1       2   46    7
15      1       1   42    8
16      1       2   54    8
17      2       1   34    9
18      2       2   16    9
19      2       1   40   10
20      2       2   36   10
21      2       1   34   11
22      2       2   16   11
23      2       1   36   12
24      2       2   18   12
25      2       1   38   13
26      2       2   32   13
27      2       1   32   14
28      2       2   14   14
29      2       1   44   15
30      2       2   20   15
31      2       1   50   16
32      2       2   43   16
33      2       1   60   17
34      2       2   45   17
35      2       1   63   18
36      2       2   67   18
37      2       1   50   19
38      2       2   36   19
39      2       1   42   20
40      2       2   34   20
41      2       1   43   21
42      2       2   32   21
```


subject
outcome
time
60
50
40
30
20
1.0
1.2
1.4
1.6
1.8
2.0

```
xyplot(outcome ~ time|method, groups = subject, type = c("g", "p","r"),index.cond=function(x,y)
{coef(lm(y ~ x))[1]},data = bk, col = c("black"))
```



```
xyplot(outcome ~ time, groups = method, type = c("r"),index.cond=function(x,y) {coef(lm(y ~ x))[1]},data = bk)
```



```
    Update of BK repeated measures analysis

R version 2.14.1 (2011-12-22)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-pc-mingw32/x64 (64-bit)

> library(lme4)


> #note brogkutlong restarts subject numbering at 1 for each method; brogkutlong2 numbe

> bk  = read.table(file="http://www-stat.stanford.edu/~rag/stat222/brogkutlong2.dat", h

> attach(bk)
> bklist = lmList(outcome ~ time|subject, data = bk) # getting difference scores the ha
> bklist
Call: lmList(formula = outcome ~ time | subject, data = bk)
Coefficients:
   (Intercept) time
1           54   -3
2           15   20
3           72   -6
4           45   -5
5           42   -3
6           49   -3
7           58   -6
8           30   12
9           52  -18
10          44   -4
11          52  -18
12          54  -18
13          44   -6
14          50  -18
15          68  -24
16          57   -7
17          75  -15
18          59    4
19          64  -14
20          50   -8
21          54  -11

Error in pooledSD(object) :
  No degrees of freedom for estimating std. dev.
# if you want the "intercept" to be level at time=1 (pretest) the
> t1 = time - 1
> bklist1 = lmList(outcome ~ t1|subject, data = bk)

> library(lattice) # make a plot for individual subjects
> xyplot(outcome ~ time|subject, groups = method, type = c("p","r"), data = bk)

# the repeated measures anova, shown in previous analysis
> bkrepaov1 = aov(outcome ~ as.factor(time)*as.factor(method)+ Error(as.factor(subject)
> summary(bkrepaov1)

Error: as.factor(subject)
                  Df Sum Sq Mean Sq F value Pr(>F)
as.factor(method)  1    847   847.5   3.627 0.0721 .
Residuals         19   4440   233.7
---
```

Brogan-Kutner lmer

STAT 222

```
BK Analysis handout          # recent version of lme4 objects to two-wave data. Rerun 10/19/17
> library(lme4)
> bk  = read.table(file="http://statweb.stanford.edu/~rag/stat222/brogkutlong2.dat", header = T)
> bk$t1 = bk$time - 1      > bk$G = bk$method - 1      > head(bk)
  method time outcome subject t1 G
1      1    1      51       1  0 0
> bklist = lmList(outcome ~ t1|subject, data = bk) # getting difference scores the hard way
> bklmera = lmer(outcome ~ t1 + t1:as.factor(method) + (t1|subject), data = bk)
Error: number of observations (=42) <= number of random effects (=42) for term (t1 | subject); the ranc
> # fix it by 'no 2-wave worries'
> bklmera = lmer(outcome ~ t1 + t1:as.factor(method) + (t1|subject), data = bk,
                                        control = lmerControl(check.nobs.vs.nRE = "warning"))
Warning messages:
1: number of observations (=42) <= number of random effects (=42) for term (t1 | subject); the random-e
> summary(bklmera)
Linear mixed model fit by REML ['lmerMod']
Formula: outcome ~ t1 + t1:as.factor(method) + (t1 | subject)  Data: bk
Control: lmerControl(check.nobs.vs.nRE = "warning")
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 subject  (Intercept) 66.45    8.152
          t1          17.31    4.161    0.87
 Residual             27.20    5.215
Number of obs: 42, groups:  subject, 21
Fixed effects:
                    Estimate Std. Error t value
(Intercept)          44.6190     2.1117  21.129
t1                    0.7057     2.9931   0.236
t1:as.factor(method)2 -12.7553     3.8035  -3.354   #method 2 is old method

> anova(bklmera) # put fixed effects in SS metric
Analysis of Variance Table
                    Df Sum Sq Mean Sq F value
t1                   1 411.79  411.79  15.142
t1:as.factor(method) 1 305.84  305.84  11.246

> confint(bklmera)
Computing profile confidence intervals ...
> # properly bombs on random effects because fitting line to 2 points |subject
> confint(bklmera, method = "boot", nsim = 1000, boot.type = "perc")
Computing bootstrap confidence intervals ...
                                  2.5 %    97.5 %
sd_(Intercept)|subject         4.9016210 11.555923
cor_t1.(Intercept)|subject    -0.3783355  1.000000
sd_t1|subject                  0.6097805  7.555747
sigma                          3.4510062  6.530776
(Intercept)                   40.5845364 48.945127
t1                            -5.4157384  6.548160
t1:as.factor(method)2        -20.1664950 -5.184535
There were 50 or more warnings (use warnings() to see the first 50)
> # bootstrap gives reasonable bounds for random effects even

> # lmer 'a' does not include pretest diffs because of random assignment, can look at that
> bklmerb = lmer(outcome ~ t1 + t1*as.factor(method) + (t1|subject), data = bk,  control = lmerControl(
> anova(bklmera, bklmerb) # compare nested models
refitting model(s) with ML (instead of REML) Data: bk
Models:
bklmera: outcome ~ t1 + t1:as.factor(method) + (t1 | subject)
bklmerb: outcome ~ t1 + t1 * as.factor(method) + (t1 | subject)
        Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
bklmera  7 315.10 327.26 -150.55   301.10
bklmerb  8 316.65 330.55 -150.32   300.65 0.4516      1     0.5016
> # extended model does not help
```

Model

Level 1 $Y = \alpha_0 + \alpha_1 t1 + \varepsilon$

(α_0 = pre, α_1 = post - pre)

Level 2 $\alpha_0 = \gamma_{00} + u_0$ no diff

$\alpha_1 = \gamma_{10} + \gamma_{11}\,\text{method} + u_1$

$=$

Combined

$Y = \gamma_{00} + \gamma_{10} t1 + \gamma_{11} t1{:}\text{method} + [u_0 + u_1 + \varepsilon]$

Repeated Measures Brogan-Kutner ix p.2

D. Rocosa

model

$$X_{ijk} = \mu + \alpha_i + \Pi_{k(i)} + \beta_j + \alpha\beta_{ij} + \beta\Pi_{jk(i)} + \epsilon_{m(ijk)} \quad (3.1)$$

$j = 1, 2$ (pretest = 1, posttest = 2),

$i = 1, 2$ (group 1 = 1, group 2 = 2),

$k = 1, 2, \ldots, n_i$, $m = 1$,

where X_{ijk} is the observed value of subject k within group i at time j,

μ is the overall mean,

α_i is the effect of group i,

$\Pi_{k(i)}$ is the effect of subject k nested within group i,

β_j is the effect of the repeated-measures variable j (*i.e.*, pretest and posttest),

$\alpha\beta_{ij}$ is the interaction of group i with level j of the repeated measures factor,

$\beta\Pi_{jk(i)}$ is the interaction of subject k within group i with level j of the repeated-measures factor.

2. Repeated-Measures Analysis of Variance for Maximal Rate of Urea Synthesis Level

Source of Variation	df	Sum of Squares	Mean Squares	F Ratio
Between Subjects	20 (n − 1)			
Groups	1	847.48	847.48 (MS_G)	3.63 (MS_G/MS_E)
Subjects Within Groups	19 (n − 2)	4440.00	233.68 (MS_E)	
Within Groups	21 (n)			
Pre/Post	1	317.69	317.69 (MS_P)	8.86 (MS_P/MS_{PE})
Groups x Pre/Post	1	407.41	407.4 (MS_{GP})	11.36 (MS_{GP}/MS_{PE})
(Pre/Post) x Subjects Within Groups	19 (n − 2)	681.21	35.85 (MS_{PE})	

Did the groups change differentially?

SAS or minitab does it

(R has problem w/ imbalance anova /error TBD

```
proc glm data=brogk;
class grp;
model m1--m2 = grp /nouni;
repeated Time 2 (1 2) / summary printe;
run;
```

OUTPUT (selected)

```
        The SAS System                16:13 Tuesday, May 16, 2000   35

                                The GLM Procedure
                        Repeated Measures Analysis of Variance
                    Tests of Hypotheses for Between Subjects Effects

Source        DF      Type III SS      Mean Square    F Value    Pr > F

grp            1      847.476190       847.476190       3.63    0.0721
Error         19     4440.000000       233.684211

                                The GLM Procedure
                        Repeated Measures Analysis of Variance
                  Univariate Tests of Hypotheses for Within Subject Effects

Source          DF      Type III SS      Mean Square    F Value    Pr > F

Time             1      317.6932234      317.6932234       8.86    0.0078
Time*grp         1      407.4075092      407.4075092      11.36    0.0032
Error(Time)     19      681.2115385       35.8532389
```

```
# Brogan-Kutner Data see http://www-stat.stanford.edu/~rag/ed351longit/brogkut.dat

# Cell means
> tapply(urea, list(method, prepost), mean)
         1        2
1 46.37500 47.12500
2 43.53846 31.46154
```

```
# Recreate repeated measures anova (nesting)
# within-groups anova to obtain the 2 error terms

#within group 1 subjXtime
> bkrepaovW1 = aov(urea[method == "1"] ~ as.factor(prepost[method == "1"])*as.factor(subj[method == "1"]))
> summary(bkrepaovW1)
                                                          Df Sum Sq Mean Sq
as.factor(prepost[method == "1"])                          1   2.25    2.25
as.factor(subj[method == "1"])                             7 915.00  130.71
                        piece of subjects within groups  Between subjects error term
as.factor(prepost[method == "1"]):as.factor(subj[method == "1"])  7 331.75   47.39
   piece of subjectsxrepeated measure  within group interaction   Within subjects error term

#within group 2 subjXtime
> bkrepaovW2 = aov(urea[method == "2"] ~ as.factor(prepost[method == "2"])*as.factor(subj[method == "2"]))
> summary(bkrepaovW2)
                                                          Df Sum Sq Mean Sq
as.factor(prepost[method == "2"])                          1  948.0   948.0
as.factor(subj[method == "2"])                            12 3525.0   293.7
                       piece of subjects within groups  Between subjects error term
as.factor(prepost[method == "2"]):as.factor(subj[method == "2"]) 12  349.5    29.1
   piece of subjectsxrepeated measure  within group interaction   Within subjects error term

#              915 + 3525 = 4440 (and 7 + 12 = 19df)   Between subjects SS error term
#              331.7 + 349.5 = 681.2 (and 7 + 12 = 19df) Within subjects SS error term

# ignore within-subjects, get
> bkrepaovBase = aov(urea ~ as.factor(prepost)*as.factor(method))
> summary(bkrepaovBase)
                                  Df Sum Sq Mean Sq F value  Pr(>F)
as.factor(prepost)                 1  542.9   542.9  4.0282 0.05190 .  #repeated measure (Within subj part)
as.factor(method)                  1  847.5   847.5  6.2884 0.01654 *  #Group  (Between subjects part)
as.factor(prepost):as.factor(method)  1  407.4   407.4  3.0230 0.09019 .  #GroupxRepeated measure Interaction
Residuals                         38 5121.2   134.8                          (Within subjects part)
```

```
# Brogan-Kutner Section 5 Equivalences
   # Groups, pooling over occasion
> sumtime = pre + post
> t.test(sumtime ~ as.factor(method), var.equal = TRUE)
    Two Sample t-test data:  sumtime by as.factor(method)
t = 1.9044, df = 19, p-value = 0.07212
95 percent confidence interval: -1.832786 38.832786
mean in group 1 mean in group 2
           93.5            75.0
> 1.904^2  [1] 3.625216  # matches F-stat for Groups (bet subj)

> imp = post - pre
> t.test(imp ~ as.factor(method), var.equal = TRUE)
  Two Sample t-test data:  imp by as.factor(method)
t = 3.3709, df = 19, p-value = 0.003209
95 percent confidence interval:  4.862645 20.791201
mean in group 1 mean in group 2
        0.75000       -12.07692
> 3.3709^2  [1] 11.36297  # matches F-stat for Groups X prepost

> t.test(imp)
 One Sample t-test data:  imp
t = -3.1581, df = 20, p-value = 0.004947
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval: -11.939835  -2.441117
mean of x -7.190476
> 3.1581^2 [1] 9.973596  # equiv to prepost, no differential change
                            BK p.232
```

```
> bksubj
   pre post method
1   51   48      1
2   35   55      1
3   66   60      1
4   40   35      1
5   39   36      1
6   46   43      1
7   52   46      1
8   42   54      1
9   34   16      2
10  40   36      2
11  34   16      2
12  36   18      2
13  38   32      2
14  32   14      2
15  44   20      2
16  50   43      2
17  60   45      2
18  63   67      2
19  50   36      2
20  42   34      2
21  43   32      2
```

```
> bkrepaov1 = aov(urea ~ as.factor(prepost)*as.factor(method)+ Error(as.factor(subj)))
> summary(bkrepaov1)
Error: as.factor(subj)
                  Df Sum Sq Mean Sq F value  Pr(>F)
as.factor(method)  1  847.5   847.5  3.6266 0.07212 .
Residuals         19 4440.0   233.7
---
Error: Within
                                  Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(prepost)                 1 542.88  542.88  15.142 0.0009823 ***Type III SS(prepost) = 317
as.factor(prepost):as.factor(method)  1 407.41  407.41  11.363 0.0032085 **
Residuals                         19 681.21   35.85
```

Repeated Measures Anova more BK

Stat 209 [redacted]

```
# Brogan-Kutner Data see http://www-stat.stanford.edu/~rag/ed351longit/brogkut.dat

# Cell means
> tapply(urea, list(method, prepost), mean)
         1        2
1 46.37500 47.12500
2 43.53846 31.46154
```

cf main BK handout

data observations as rows (42)

```
# Recreate repeated measures anova (nesting)
# within-groups anova to obtain the 2 error terms
```

Do repeated measures anova by crossed designs on subjects.

```
#within group 1 subjXtime
> bkrepaovW1 = aov(urea[method == "1"] ~ as.factor(prepost[method == "1"])*as.factor(subj[method == "1"]))
> summary(bkrepaovW1)

                                                                  Df Sum Sq Mean Sq
as.factor(prepost[method == "1"])                                  1   2.25    2.25
as.factor(subj[method == "1"])                                     7 915.00  130.71
                          piece of subjects within groups  Between subjects error term
as.factor(prepost[method == "1"]):as.factor(subj[method == "1"])   7 331.75   47.39
   piece of subjectsxrepeated measure  within group interaction   Within subjects error term

#within group 2 subjXtime
> bkrepaovW2 = aov(urea[method == "2"] ~ as.factor(prepost[method == "2"])*as.factor(subj[method == "2"]))
> summary(bkrepaovW2)

                                                                  Df Sum Sq Mean Sq
as.factor(prepost[method == "2"])                                  1  948.0   948.0
as.factor(subj[method == "2"])                                    12 3525.0   293.7
                          piece of subjects within groups  Between subjects error term
as.factor(prepost[method == "2"]):as.factor(subj[method == "2"]) 12  349.5    29.1
   piece of subjectsxrepeated measure  within group interaction   Within subjects error term

#               915 + 3525 = 4440 (and 7 + 12 = 19df)   Between subjects SS error term
#               331.7 + 349.5 = 681.2 (and 7 + 12 = 19df) Within subjects SS error term

# ignore within-subjects, get
> bkrepaovBase = aov(urea ~ as.factor(prepost)*as.factor(method))
> summary(bkrepaovBase)
                                 Df Sum Sq Mean Sq F value  Pr(>F)
as.factor(prepost)                1  542.9   542.9  4.0282 0.05190 .  #repeated measure (Within subj part)
as.factor(method)                 1  847.5   847.5  6.2884 0.01654 *  #Group  (Between subjects part)
as.factor(prepost):as.factor(method) 1 407.4  407.4  3.0230 0.09019 .  #GroupxRepeated measure Interaction
Residuals                        38 5121.2   134.8                      (Within subjects part)
```

```
# Brogan-Kutner Section 5 Equivalences
   # Groups, pooling over occasion
> sumtime = pre + post
> t.test(sumtime ~ as.factor(method), var.equal = TRUE)
    Two Sample t-test data:  sumtime by as.factor(method)
t = 1.9044, df = 19, p-value = 0.07212
95 percent confidence interval: -1.832786 38.832786
mean in group 1 mean in group 2
          93.5            75.0
> 1.904^2  [1] 3.625216  # matches F-stat for Groups (bet subj)

> imp = post - pre
> t.test(imp ~ as.factor(method), var.equal = TRUE)
  Two Sample t-test data:  imp by as.factor(method)
t = 3.3709, df = 19, p-value = 0.003209
95 percent confidence interval:  4.862645 20.791201
mean in group 1 mean in group 2
        0.75000       -12.07692
> 3.3709^2  [1] 11.36297  # matches F-stat for Groups X prepost
```

main event differential change by t-test

```
> t.test(imp)
 One Sample t-test data:  imp
t = -3.1581, df = 20, p-value = 0.004947
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval: -11.939835  -2.441117
mean of x -7.190476
> 3.1581^2 [1] 9.973596  # equiv to prepost, no differential change
                          BK p.232
```

subj as rows format

```
> bksubj
   pre post method
1   51   48      1
2   35   55      1
3   66   60      1
4   40   35      1
5   39   36      1
6   46   43      1
7   52   46      1
8   42   54      1
9   34   16      2
10  40   36      2
11  34   16      2
12  36   18      2
13  38   32      2
14  32   14      2
15  44   20      2
16  50   43      2
17  60   45      2
18  63   67      2
19  50   36      2
20  42   34      2
21  43   32      2
```

```
> bkrepaov1 = aov(urea ~ as.factor(prepost)*as.factor(method)+ Error(as.factor(subj)))
> summary(bkrepaov1)
Error: as.factor(subj)
                  Df Sum Sq Mean Sq F value  Pr(>F)
as.factor(method)  1  847.5   847.5  3.6266 0.07212 .
Residuals         19 4440.0   233.7
---
Error: Within
                                    Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(prepost)                   1 542.88  542.88  15.142 0.0009823 ***
as.factor(prepost):as.factor(method)  1 407.41  407.41  11.363 0.0032085 **
Residuals                            19 681.21   35.85
```

subj x prepost x method

R does the repeated meas design See Baron+Li

Sequential SS issue w/ prepost SS

[redacted] BK lmer

```
    Update of BK repeated measures analysis
> library(lme4)
> #note brogkutlong restarts subject numbering at 1 for each method; brogkutlong2 numbers sequentially
> bk  = read.table(file="http://www-stat.stanford.edu/~rag/stat222/brogkutlong2.dat", header = T)
> attach(bk)
> bklist = lmList(outcome ~ time|subject, data = bk) # getting difference scores the hard way
> bklist
Call: lmList(formula = outcome ~ time | subject, data = bk)
Coefficients:
   (Intercept) time
1           54   -3
2           15   20
...
21          54  -11
# if you want the "intercept" to be level at time=1 (pretest) the
> t1 = time - 1
> bklist1 = lmList(outcome ~ t1|subject, data = bk)

> library(lattice) # make a plot for individual subjects
> xyplot(outcome ~ time|subject, groups = method, type = c("p","r"), data = bk)

# the repeated measures anova, shown in previous analysis
> bkrepaov1 = aov(outcome ~ as.factor(time)*as.factor(method)+ Error(as.factor(subject)))
> summary(bkrepaov1)
Error: as.factor(subject)
                  Df Sum Sq Mean Sq F value Pr(>F)
as.factor(method)  1    847   847.5   3.627 0.0721 .
Residuals         19   4440   233.7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Error: Within
                                 Df Sum Sq Mean Sq F value   Pr(>F)
as.factor(time)                   1  542.9   542.9   15.14 0.000982 ***
as.factor(time):as.factor(method)  1  407.4   407.4   11.36 0.003209 **
Residuals                         19  681.2    35.9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# as noted R does Type I SS, Type III SS for time is 317 (SAS etc); interaction is prime concern,
   that (407) matches SAS PROC GLM
```

truncated

best do bk$t1 = bk$time - 1

better version

fun plots

unequal group sizes makes non-orthog design

bigger than Type I

matches SAS, publication

```
#so let's try an lmer model: level 1 outcome ~ time; level 2 slope (diff score) depends on method
> bklmera = lmer(outcome ~ I(time - 1) + I(time-1):as.factor(method) + (time|subject), data = bk)
> summary(bklmera)
Linear mixed model fit by REML
Formula: outcome ~ I(time - 1) + I(time - 1):as.factor(method) + (time |      subject)
   Data: bk
   AIC   BIC logLik deviance REMLdev
 305.7 317.9 -145.9    301.1   291.7
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 subject  (Intercept) 35.000   5.9161
          time        21.455   4.6320   0.220
 Residual             25.125   5.0124
Number of obs: 42, groups: subject, 21
Fixed effects:
                                Estimate Std. Error t value
(Intercept)                       44.619      2.112  21.130
I(time - 1)                       -5.672      1.902  -2.981
I(time - 1):as.factor(method)1     6.378      1.902   3.354

Correlation of Fixed Effects:
            (Intr) I(t-1)
I(time - 1) 0.028
I(-1):s.()1 0.000  0.238
# so interaction matches F-statistic from repeated measures anova
> 3.354^2
[1] 11.24932
# AND lmer gets the occasions (time) term "correct" in the test statistic
> 2.981^2
[1] 8.886361
# this matches F-statistic in publication (and SAS) repeated measures output of 8.86 for pre/post (time)
# whereas the aov above has F-statistic 15.1
# SS not comparable with anova because here were are modeling level 1 params, not outcome
So before looking at other small details, let us declare an lmer victory over non-othogonal designs
```

lmer rules

Type II SAS

Model

Level 1 within subject

t1 = time - 1 better this way

$y = \alpha_0 + \alpha_1 t1 + \varepsilon$

α_0 = pre

α_1 = post - pre

Level 2

$\alpha_0 = \gamma_{00} + u_0$

$\alpha_1 = \gamma_{10} + \gamma_{11}\,\text{method} + u_1$

Combined

$y = \gamma_{00} + \gamma_{10} t1 + \gamma_{11} t1{:}\text{method} + [u_0 + u_1 t1 + \varepsilon]$

extended version posted bklmer

lmer 2-wave

The most recent version of lme4 (not the one I've been using from 2014) objects to two-wave data.
I confirmed this by starting a new fully updated R-version with a newly downloaded `lme4`, which for the Brogan-Kutner example

```
> bk  = read.table(file="http://statweb.stanford.edu/~rag/stat222/brogkutlong2.dat", header = T)
> bklist = lmList(outcome ~ time|subject, data = bk) # getting difference scores the hard way
> bk$t1 = bk$time - 1
> bklmera = lmer(outcome ~ t1 + t1:as.factor(method) + (t1|subject), data = bk)
Error: number of observations (=42) <= number of random effects (=42) for term (t1 | subject); the random-effects parameters and the residual variance (or scale parameter) are probably unidentifiable
```

A help thread that indicated appending `control = lmerControl(check.nobs.vs.nRE = "warning")` in the lmer statement will get you an functional lmer object that you can do summary on and get fixed effects. Random effects and CI for such appear not to work well
`https://github.com/lme4/lme4/issues/175`

The work-around I suggest is to employ the older brother of `lme4`, package `nlme`, function `lme` for two-wave data. The nlme package is part of base R and is still widely used (in fact the brand new book 'Multilevel models with R' annoyingly uses nlme as the primary).
We met package `nlme` briefly in week 9, as the Joint Models package uses nlme for the measured variables (time trajectories) portion of the analysis.
The code above changes to (notice the clunkier syntax for the random part of the mixed-model).

```
> bklmea = lme(outcome ~ t1 + t1:as.factor(method), random = ~ t1|subject, data = bk)
> summary(bklmea)
```

A short session using `lme` for the Brogan-Kutner data is provided here

```
> bklmera = lmer(outcome ~ t1 + t1:as.factor(method) + (t1|subject), data = bk)
Error: number of observations (=42) <= number of random effects (=42) for term (t1 | subject); the random-effects paramete

> # fix it by 'no 2-wave worries'
> bklmera = lmer(outcome ~ t1 + t1:as.factor(method) + (t1|subject), data = bk,  control = lmerControl(check.nobs.vs.nRE =
Warning messages:
1: number of observations (=42) <= number of random effects (=42) for term (t1 | subject); the random-effects parameters a
2: In checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv,  :
  Model is nearly unidentifiable: large eigenvalue ratio
 - Rescale variables?

> summary(bklmera)
Linear mixed model fit by REML ['lmerMod']
Formula: outcome ~ t1 + t1:as.factor(method) + (t1 | subject)
   Data: bk
Control: lmerControl(check.nobs.vs.nRE = "warning")

REML criterion at convergence: 290.3

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.9936 -0.4127 -0.1596  0.4288  1.7313

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 subject  (Intercept) 66.45    8.152
          t1          17.31    4.161    0.87
 Residual             27.20    5.215
Number of obs: 42, groups:  subject, 21

Fixed effects:
                     Estimate Std. Error t value
(Intercept)           44.6190     2.1117  21.129
t1                     0.7057     2.9931   0.236
t1:as.factor(method)2 -12.7553     3.8035  -3.354

Correlation of Fixed Effects:
            (Intr) t1
t1           0.018
t1:s.fct()2  0.000 -0.787

> anova(bklmera) # put fixed effects in SS metric
Analysis of Variance Table
                    Df Sum Sq Mean Sq F value
```

```
t1                    1 411.79  411.79  15.142
t1:as.factor(method)  1 305.84  305.84  11.246

> confint(bklmera)
Computing profile confidence intervals ...
                            2.5 %    97.5 %
.sig01                   0.000000       Inf
.sig02                  -1.000000  1.000000
.sig03                   0.000000       Inf
.sigma                   0.000000       Inf
(Intercept)             40.387933 48.850163
t1                      -5.184306  6.589559
t1:as.factor(method)2 -20.263453 -5.247139
There were 50 or more warnings (use warnings() to see the first 50)
> # properly bombs on random effects because fitting line to 2 points |subject

> confint(bklmera, method = "boot", nsim = 1000, boot.type = "perc")
Computing bootstrap confidence intervals ...
                                  2.5 %    97.5 %
sd_(Intercept)|subject       4.9016210 11.555923
cor_t1.(Intercept)|subject  -0.3783355  1.000000
sd_t1|subject                0.6097805  7.555747
sigma                        3.4510062  6.530776
(Intercept)                 40.5845364 48.945127
t1                          -5.4157384  6.548160
t1:as.factor(method)2      -20.1664950 -5.184535
There were 50 or more warnings (use warnings() to see the first 50)
> # bootstrap gives reasonable bounds for random effects even

> # lmer 'a' does not include pretest diffs because of random assignment, can look at that
```

```
MTB > read 'a:\351\brogkut.dat' c1-c4
Entering data from file: a:\351\brogkut.dat
     42 rows read.
MTB > name c1 'method'
MTB > name c2 'prepost'
MTB > name c3 'outcome'
MTB > name c4 'subject'
MTB > info

Column   Name          Count
C1       method           42
C2       prepost          42
C3       outcome          42
C4       subject          42

REPEATED MEASURES ANALYSIS

MTB > glm outcome = subject(method) + method|prepost;
SUBC> random subject;
SUBC> ems;
SUBC> means method|prepost.

General Linear Model


Factor            Type Levels Values
subject(method) random     21  1  2  3  4  5  6  7  8  1  2  3  4  5  6  7  8
                                9 10 11 12 13
method           fixed      2 1 2
prepost          fixed      2 1 2

Analysis of Variance for outcome, using Adjusted SS for Tests

Source            DF     Seq SS     Adj SS     Adj MS       F      P
subject(method)   19    4440.00    4440.00     233.68    6.52  0.000
method             1     847.48     847.48     847.48    3.63  0.072
prepost            1     542.88     317.69     317.69    8.86  0.008
method*prepost     1     407.41     407.41     407.41   11.36  0.003
Error             19     681.21     681.21      35.85
Total             41    6918.98

Unusual Observations for outcome

Obs   outcome       Fit   StDev Fit  Residual   St Resid
  3   35.0000   44.6250      4.4908   -9.6250     -2.43R
  4   55.0000   45.3750      4.4908    9.6250      2.43R

R denotes an observation with a large standardized residual.

Expected Mean Squares, using Adjusted SS

Source              Expected Mean Square for Each Term
 1 subject(method)  (5) +  2.0000(1)
 2 method           (5) +  2.0000(1) + Q[2, 4]
 3 prepost          (5) + Q[3, 4]
 4 method*prepost   (5) + Q[4]
 5 Error            (5)

Error Terms for Tests, using Adjusted SS

Source              Error DF  Error MS  Synthesis of Error MS
 1 subject(method)     19.00     35.85  (5)
 2 method              19.00    233.68  (1)
 3 prepost             19.00     35.85  (5)
 4 method*prepost      19.00     35.85  (5)

Variance Components, using Adjusted SS

Source           Estimated Value
```

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

Error: Within
                                 Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(time)                   1  542.9   542.9   15.14 0.000982 ***
as.factor(time):as.factor(method) 1  407.4   407.4   11.36 0.003209 **
Residuals                        19  681.2    35.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# as noted R does Type I SS, Type III SS for time is 317 (SAS etc); interaction is prim
   that (407) matches SAS PROC GLM
```

```
#so let's try an lmer model: level 1 outcome ~ time; level 2 slope (diff score) depends

> bklmera = lmer(outcome ~ I(time - 1) + I(time-1):as.factor(method) + (time|subject),
> summary(bklmera)
Linear mixed model fit by REML
Formula: outcome ~ I(time - 1) + I(time - 1):as.factor(method) + (time |      subject)
   Data: bk
   AIC   BIC logLik deviance REMLdev
 305.7 317.9 -145.9    301.1   291.7
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 subject  (Intercept) 35.000   5.9161
          time        21.455   4.6320   0.220
 Residual             25.125   5.0124
Number of obs: 42, groups: subject, 21

Fixed effects:
                               Estimate Std. Error t value
(Intercept)                      44.619      2.112  21.130
I(time - 1)                      -5.672      1.902  -2.981
I(time - 1):as.factor(method)1    6.378      1.902   3.354

Correlation of Fixed Effects:
            (Intr) I(t-1)
I(time - 1) 0.028
I(-1):s.()1 0.000  0.238
```

```
> anova(bklmera)
Analysis of Variance Table
                             Df Sum Sq Mean Sq F value
I(time - 1)                   1 380.69  380.69  15.152
I(time - 1):as.factor(method) 1 282.54  282.54  11.246

# so interaction matches F-statistic from repeated measures anova
> 3.354^2
[1] 11.24932
# AND lmer gets the occasions (time) term "correct" in the test statistic
> 2.981^2
[1] 8.886361
# this matches F-statistic in publication (and SAS) repeated measures output of 8.86 fo
# whereas the aov above has F-statistic 15.1
# SS not comparable with anova because here were are modeling level 1 params, not outco

So before looking at other small details, let us declare an lmer victory over non-othog
```

```
# even if you let method be numerical (1,2) inadvertently it works ok here
> bklmer = lmer(outcome ~ I(time - 1) + I(time-1):method + (time|subject), data = bk)
> bklmer
Linear mixed model fit by REML
Formula: outcome ~ I(time - 1) + I(time - 1):method + (time | subject)
   Data: bk
   AIC   BIC logLik deviance REMLdev
 304.3 316.5 -145.2    301.1   290.3
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 subject  (Intercept) 35.000   5.9161
          time        21.455   4.6320   0.220
 Residual             25.125   5.0124
Number of obs: 42, groups: subject, 21

Fixed effects:
                   Estimate Std. Error t value
(Intercept)          44.619      2.112  21.130
I(time - 1)          13.461      6.429   2.094
I(time - 1):method  -12.755      3.804  -3.354

Correlation of Fixed Effects:
            (Intr) I(t-1)
I(time - 1)  0.008
I(tm-1):mth  0.000 -0.958
> anova(bklmer)
Analysis of Variance Table
                   Df Sum Sq Mean Sq F value
I(time - 1)         1 380.69  380.69  15.152
I(time - 1):method  1 282.54  282.54  11.246
```

```
# more general model also lets intercept (time1) differ by method, but randomization sh
> bklmer2a = lmer(outcome ~ I(time - 1)*as.factor(method) + (time|subject), data = bk)
> bklmer2a
Linear mixed model fit by REML
Formula: outcome ~ I(time - 1) * as.factor(method) + (time | subject)
   Data: bk
   AIC BIC logLik deviance REMLdev
 301.1 315 -142.6    300.6   285.1
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 subject  (Intercept) 36.693   6.0575
          time        21.058   4.5889   0.241
 Residual             25.324   5.0323
Number of obs: 42, groups: subject, 21

Fixed effects:
                                Estimate Std. Error t value
(Intercept)                       46.375      3.473  13.354
I(time - 1)                        0.750      2.994   0.251
as.factor(method)2                -2.837      4.414  -0.643
I(time - 1):as.factor(method)2   -12.827      3.805  -3.371

Correlation of Fixed Effects:
            (Intr) I(t-1) as.()2
I(time - 1)  0.029
as.fctr(m)2 -0.787 -0.023
I(-1):s.()2 -0.023 -0.787  0.029
> anova(bklmer2a)
```

```
Analysis of Variance Table
                             Df Sum Sq Mean Sq F value
I(time - 1)                   1 383.72  383.72 15.1525
as.factor(method)             1   7.47    7.47  0.2952
I(time - 1):as.factor(method)  1 287.73  287.73 11.3621
```

```
> summary(bklmer2a)
Linear mixed model fit by REML
Formula: outcome ~ I(time - 1) * as.factor(method) + (time | subject)
   Data: bk
   AIC BIC logLik deviance REMLdev
 301.1 315 -142.6    300.6   285.1
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 subject  (Intercept) 36.693   6.0575
          time        21.058   4.5889   0.241
 Residual             25.324   5.0323
Number of obs: 42, groups: subject, 21

Fixed effects:
                             Estimate Std. Error t value
(Intercept)                    46.375      3.473  13.354
I(time - 1)                     0.750      2.994   0.251
as.factor(method)2             -2.837      4.414  -0.643
I(time - 1):as.factor(method)2 -12.827      3.805  -3.371

Correlation of Fixed Effects:
            (Intr) I(t-1) as.()2
I(time - 1)  0.029
as.fctr(m)2 -0.787 -0.023
I(-1):s.()2 -0.023 -0.787  0.029
```

```
> anova(bklmer, bklmer2a) # the extra method main effect here doesn't help
Data: bk
Models:
bklmer: outcome ~ I(time - 1) + I(time - 1):method + (time | subject)
bklmer2a: outcome ~ I(time - 1) * as.factor(method) + (time | subject)
         Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
bklmer    7 315.11 327.28 -150.56
bklmer2a  8 316.65 330.55 -150.32 0.4654      1     0.4951
```

```
> install.packages("ez") # I tried the "ez" package, but didn't help with anova
```

Another approach to BK; pretest as covariate or t-test on posttest; see review quests

Stat 209 [redacted]

ANCOVA CNRL equations

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

$G = 0, 1$ group membership

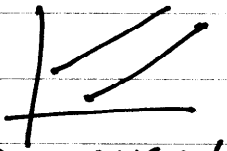
$$Y = \gamma_0 + \gamma_1 G + \gamma_2 X$$

ancova

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

constant treatment effect

$\hat{\gamma}_2 = \hat{\delta}_p$ ave within groups slopes



---

more general model (CNRL)

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

"interaction" term