Statistics 222, Education 351A Autumn 2020

Statistical Methods for Longitudinal Research

Autumn 2020 Remote Asynchronous Instruction

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

To see full course materials from Autumn 2018 go here

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

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

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

Preliminary Course Outline

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

Texts and Resources for Course Content

1. Garrett M. Fitzmaurice Nan M. Laird James H. Ware Applied Longitudinal Analysis (Wiley Series in Probability and Statistics; 2nd ed 2011)
Text Website second edition website Text lecture slides
2. Judith D. Singer and John B. Willett . Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence New York: Oxford University Press, March, 2003.
Text web page Text data examples at UCLA IDRE Powerpoint presentations good gentle intro to modelling collections of growth curves (and survival analysis) is Willett and Singer (1998)
3. Douglas M. Bates. lme4: Mixed-effects modeling with R February 17, 2010 Springer (chapters). A merged version of Bates book: lme4: Mixed-effects modeling with R January 11, 2010 has been refound Manual for R-package lme4 and mlmRev, Bates-Pinheiro book datasets.
Additional Doug Bates materials. Collection of all Doug Bates lme4 talks Mixed models in R using the lme4 package Part 2: Longitudinal data, modeling interactions Douglas Bates 8th International Amsterdam Conference on Multilevel Analysis 2011-03-16 another version

Original Bates-Pinheiro text (2000). Mixed-Effects Models in S and S-PLUS (Stanford access). Appendix C has non-linear regression models.
Fitting linear mixed-effects models using lme4, *Journal of Statistical Software* Douglas Bates Martin Machler Ben Bolker. Technical topics: Mixed models in R using the lme4 package Part 4: Theory of linear mixed models
4. A handbook of statistical analyses using R (second edition). Brian Everitt, Torsten Hothorn CRC Press, Index of book chapters Stanford access Longitudinal chapters: Chap11 Chap12 Chap13. Data sets etc Package 'HSAUR2' August 2014, Title A Handbook of Statistical Analyses Using R (2nd Edition)
There is now a third edition of HSAUR, but full text not yet available in crcnetbase.com. CRAN HSAUR3 page with Vignettes (chapter pieces) and data in reference manual
5. Peter Diggle , Patrick Heagerty, Kung-Yee Liang , Scott Zeger. Analysis of Longitudinal Data 2nd Ed, 2002
Amazon page Peter Diggle home page Book data sets
A Short Course in Longitudinal Data Analysis Peter J Diggle, Nicola Reeve, Michelle Stanton (School of Health and Medicine, Lancaster University), June 2011 earlier version associated exercises: Lab 1 Lab2 Lab3
6. Longitudinal and Panel Data: Analysis and Applications for the Social Sciences by Edward W. Frees (2004). Full book available and book data and programs (mostly SAS).
7. Growth Curve Analysis and Visualization Using R. Daniel Mirman Chapman and Hall/CRC 2014 Print ISBN: 978-1-4665-8432-7 Stanford Access Mirman web page (including data links).
8. *Longitudinal Data Analysis* Edited by Geert Verbeke , Marie Davidian , Garrett Fitzmaurice , and Geert Molenberghs Chapman and Hall/CRC 2008. online supplement for LDA book .
9. Verbeke, G. and Molenberghs, G. (2000). Linear Mixed Models for Longitudinal Data. Springer Series in Statistics. New-York: Springer. Extended presentation: Introduction to Longitudinal Data Analysis
A shorter exposition: Methods for Analyzing Continuous, Discrete, and Incomplete Longitudinal Data
10. Survival analysis Rupert G. Miller. Available as Stanford Tech Report
11. Event History Analysis with R (Stanford access). Goran Brostrom CRC Press 2012. R-package `eha`
12. John D. Kalbfleisch , Ross L. Prentice The Statistical Analysis of Failure Time Data 2nd Ed
Amazon page online from Wiley
13. Klein J, Moeschberger M (2003). Survival Analysis, 2nd edition. New York: Springer.
14. Therneau TM, Grambsch PM (2000). Modeling Survival Data: Extending the Cox Model. New York: Springer.
15. Advanced survival analysis topics.
Interval-Censored Time-to-Event Data Methods and Applications Chapman and Hall/CRC 2012 (esp Chap 14--glrt).
Recurrent Events: Chapter 9 of Kalbfleisch and Prentice (2nd edition), "Modeling and Analysis of Recurrent Event Data".
Cook, R. J. and Lawless, J. F. (2007). The Statistical Analysis of Recurrent Events. (Stanford access) Springer, New. York.
Joint Models for Longitudinal and Time-to-Event Data. With Applications in R. Dimitris Rizopoulos. Chapman and Hall/CRC 2012(Stanford access) Book website

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

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

Course Problem Set 2020 to be posted

Cumulative Collection of Course Handounts 2020 to be posted

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

Course Content: Files, Readings, Examples

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

In the news (2018 vintage)

1. Now they tell us.... Daily aspirin may be harmful for healthy older adults, large study finds Publication: Effect of Aspirin on Disability-free Survival in the Healthy Elderly and related articles. The New England Journal of Medicine nejm.org September 2018.

2. *From 2017* Sedentary behavior can cause death (Daily Mail). Publication: Patterns of Sedentary Behavior and Mortality in U.S. Middle-Aged and Older Adults: A National Cohort Study. Ann Intern Med. 2017.

Lecture Topics

A. Longitudinal research overview

B. Examples, illustrations for longitudinal research overview, taken from course resources above:
Laird,Ware (#1) slides 1-16; Diggle (#5) slides 4-14, 22-28 Verbeke (#9) slides from Ch 2 and Sec3.3

C. Data Analysis Examples of Model Fitting for Individual Trajectories and Histories.
Motto: Individual trajectories are the proper starting point for longitudinal data analysis
ascii version of class handout annotated version pdf version with plots datasets
Starting up R-addendum: installing packages and obtaining data (sleepstudy in lme4)
Additional materials for the trajectory examples
For Count Data (glm) example. Link functions for generalized linear mixed models (GLMMs), Bates slides (pdf pages 11-18)
AIDS in Belgium example, (from Simon Wood) single trajectory, count data using glm. Rogosa R session for aids data
aditional expositions of AIDS data, Poisson regression: Duke Kentucky
A *very* comprehensive introduction to analysis of count data Regression Models for Count Data in R Achim Zeileis Christian Kleiber Simon Jackman (Stanford University)
Non-linear models, esp logistic. From week 1, also week 3 Self-Starting Logistic model SSlogis help page, do `?SSlogis` post of annotated logistic curve with SSlogis arguments
Trend in Proportions: College fund raising example prop.trend.test help page `?prop.trend.test` in R-session.
Trend in proportions, group growth, Cochran-Armitage test. Expository paper: G. Salanti and K. Ulm (2003): Tests for Trend in Binary Response (SU access)

WEEK 1 Review Questions

1. For the straight-line (constant rate of change) fit example to subj 372 in the sleepstudy data. Obtain a confidence interval for the rate of change from the OLS fit. Now compare the OLS fit with day-to-day differences. Under the constant rate of change model these 9 day to day differences also estimate the rate of change. Obtain a estimate of the mean and a confidence interval for rate of change from these first differences. Compare with OLS results.

Solution for question 1

2. Revisit the Belgium Aids data example (counts of new cases by year). Use the parameter estimates for `am2` (quadratic in time glm fit) to compute by hand (or calculator) the values of the glm fit at year = 5 and year = 9. Compare those values with results from the model `am2` using `predict`

Solution for question 2

3. Paul Rosenbaum has a little data set on growth in vocabulary that I grabbed from his Wharton coursesite. Following the *chicks* class example, plot these data and try to fit a logistic growth curve to these data. What is the estimate of the final vocab level (asymptote)? Compare the data and the fits from the logistic growth curve.
For reference, Self-Starting Logistic model SSlogis help page, do `?SSlogis` post of annotated logistic curve with SSlogis arguments additional tools in the grofit package

Solution for question 3

4. More on autocorrelation[extension/enrichment]. In standard regression courses you may have seen in addition to Durbin-Watson test for AR(1) (dwtest()), versions of the Cochrane-Orcutt procedure for remediation. Uses a first difference transformation of the data with an estimate of the autocorrelation (therefore hopeless when you have 3,4 5 observations per unit). To illustrate the statements in class and the similarities to OLS result, the solution to this problem does the straight-line and polynomial examples from the Week 1 class handout using the R-package `orcutt`

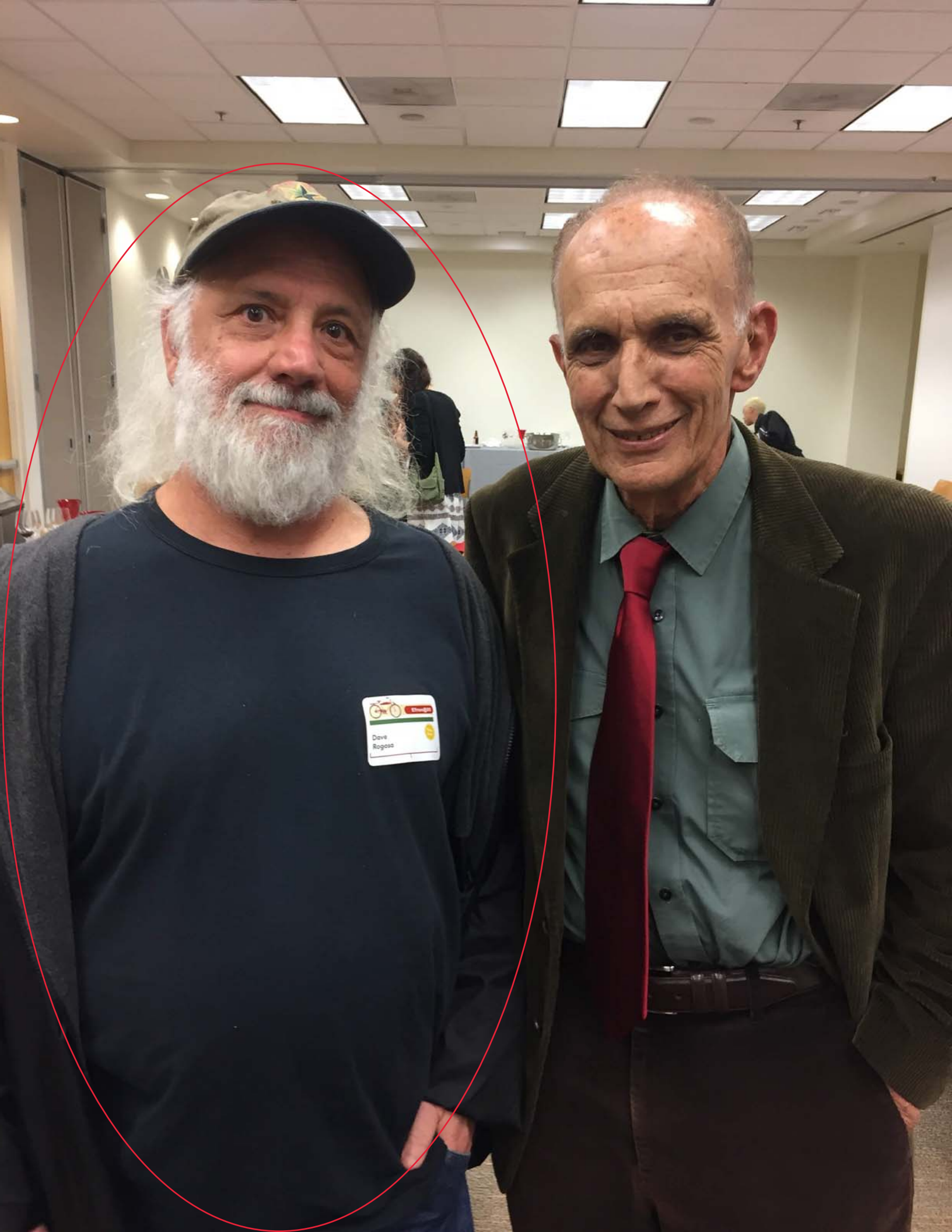
Solution for question 4

WEEK 1 Exercises

1. Straight-line fits for NC Fem data: North Carolina Achievement Data (see Williamson, Applebaum, Epanchin, 1991). These education data are eight yearly observations on achievement test scores in math (Y), for 277 females each followed from grade 1 to grade 8, with a verbal ability background measure (W).
North Carolina, female math performance (also in Rogosa-Saner) North Carolina data (wide format); NC data (long)
a. Here we will use the 8 yearly observations on female ID `705810`, which you can obtain from either the long form or wide form of these data.
For that female, what is the rate of improvement over grades 1 through 8? Compare the observed improvement for grades 1 through 8 (the *difference score*) with the amount of improvement indicated by the model fit. Obtain a 95% confidence interval for each (if possible).
b. More on OLS and the difference score. Refer to an old publication: 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; Equation 4, page 728, shows a useful form for the OLS slope. (actually reading the first three pages of that pub is a decent intro to the growth curve topic.) For equally spaced data, that Eq (4) gives a useful equivalence between difference scores (amounts of change) and OLS slopes (multiply rates of change by time interval). For the *part a* NC data show that the OLS

slope can be expressed as a weighted sum of the four differences: { 8-1,7-2,6-3,5-4}. [to say that better {score at time 8 minus score at time 1; score at time 7 minus score at time 2; ...} and so forth] Seperately, consider three observations at taken at equally spaced time intervals: What is a simple expression for the OLS slope (rate of change)?

2. Revisit the Berkeley Growth Data example from week 1 lecture. Consider the quadratic (polynomial degree 2) fit to these data, and also a (innapropriate?) constant-rate-of-change (straight-line) fit to these data. Then refer to Seigel, D. G. Several approaches for measuring average rates of change for a second degree polynomial. The American Statistician, 1975, 29, 36-37. JStor Link for equivalences for the slope of the straight-line fit to an *average rate of change* for the quadratic fit. Compare Seigel 'Approach 3" to 'Approach 1'.


Dave
Rogosa