**PHP 2690D Practical Data Analysis**

**Fall 2013**

**Meeting Times** Mon, Wed 1030-1150

**Location** Room 241 121 South Main St

**Instructor** Christopher Schmid

121 South Main St. Room 805

Phone: 3-6453

Email: [christopher\_schmid@brown.edu](mailto:christopher_schmid@brown.edu)

Office Hours: By appointment

# Description

Analyzing data is a core skill for many scientists. In addition to familiarity with probability and statistics, good data analysis requires skill in computing, data management, effective presentation of results and communication with scientific colleagues. The data analyst must be able to translate the scientific question and hypothesis into a testable form, advise about data to be collected, manipulate the data into a computable format and carry out computations to fit, test and check models.

This course is designed for graduate and advanced undergraduate students who will be analyzing data with scientific colleagues and who want to develop a practical hands-on toolkit and gain experience in distilling complex statistical information into formats understandable to colleagues. Topics including data collection and management, exploratory data analysis, fitting and checking models, simulation, handling missing data and presentation of results will be developed through a series of case studies based on different types of data requiring a variety of statistical methods. Statistical programming techniques including functions, graphs and tables will be emphasized. The R programming environment will be emphasized, although students may use other packages.

**Objectives:**

Upon completion of the course, students should be able to manipulate, program, analyze, display and present data and statistical models so that they are comprehensible for the non-statistical expert scientific collaborator.

**Prerequisites**

Students should have courses in probability and statistical inference as well as regression analysis. Familiarity with the R programming language or some other statistical programming language is advisable. Permission of the instructor is required to enroll for the course.

# Performance Evaluation

Participants will be graded based on class participation, weekly homework and a final project. To receive full credit, all assignments must be turned in when due.

# Recommended Texts

Alain F Zuur, Elena N Ieno and Erik HWG Meesters. *A Beginner’s Guide to R*. NY: Springer-Verlag, 2008. [BEG]

Phil Spector. *Data Manipulation with R*. NY: Springer-Verlag, 2008. [DATA]

Paul Murrell. *R Graphics*. NY: Chapman and Hall/CRC, 2006. [GRAPH]

Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge U Press, 2007. [ANALYSIS]

**Other Suggested Texts**

Owen Jones, Robert Maillardet and Andrew Robinson. *Introduction to Scientific Programming and Simulation Using R.* NY: CRC Press, 2009.

Hadley Wickham. *ggplot2*. NY: Springer-Verlag, 2009.

Deepayan Sarkar. *Lattice: Multivariate Data Visualization with R*. NY: Springer-Verlag, 2008.

Edward Tufte. *The Visual Display of Quantitative Information*, 2nd ed. Cheshire, CT: Graphics Press, 2001.

Norman Matloff. *The Art of R Programming*. San Francisco: no nstarch press, 2011.

Bradley Efron and Robert J Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1993. [EFRON]

**Papers**

1. McAlindon T, Formica M, Schmid CH and Fletcher J. Changes in barometric pressure and ambient temperature influence osteoarthritis pain. *American Journal of Medicine* 120:429-434, 2007. [Weather data 200 pts x 6 times x 85 variables]
2. Klempner MS, Hu LT, Evans J, Schmid CH, Johnson GM, Trevino RP, Norton D, Levy L, Wall D, McCall J, Kosinski M, and Weinstein A. Two Controlled Trials of Antibiotic Treatment in Patients with Persistent Symptoms and a History of Lyme Disease. *New England Journal of Medicine* 345: 85-92, 2001. [Lyme data 128 pts x variable times x lots of variables]
3. Beidleman BA, Tighiourart H, Schmid CH, Fulco CS and Muza SR. Predictive Models of Acute Mountain Sickness after Rapid Ascent to Various Altitudes. *Medicine and science in Sports and Exercise* 45:792-800, 2013. [Mountain sickness data 255 x variable times x 20 variables]
4. Jafar TH, Schmid CH, Landa M, Giatras I, Toto R, Remuzzi G, Maschio G, Brenner BM, Kamper A, Zucchelli P, Becker G, Himmelmann A, Bannister K, Landais P, Shahinfar S, de Jong PE, de Zeeuw D, Lau J and Levey AS for the AIPRD Study Group. Angiotensin-Converting Enzyme Inhibitors and Progression of Nondiabetic Renal Disease: A Meta-Analysis of Patient-Level Data. *Annals of Internal Medicine,* 135:73-87, 2001. [AIPRD data 1860 pts x variable time x 48 variables]
5. Wang C, Schmid CH, Hibberd PL, Kalish R, Roubenoff R, Rones R and McAlindon T. Tai Chi Is Effective in Treating Knee Osteoarthritis: A Randomized Controlled Trial. *Arthritis & Rheumatism (Arthritis Care & Research)* 61:1545–1553, 2009. [Tai Chi data 40 pts x 85 variables]
6. Schmid CH, Landa M, Jafar TH, Giatras I, Karim T, Reddy M, Stark PC and Levey AS for the Angiotensin-Converting Enzyme Inhibition in Progressive Renal Disease (AIPRD) Study Group. Constructing a Database of Individual Clinical Trials for Longitudinal Analysis. *Controlled Clinical Trials*, 24:324-340, 2003.
7. Inker LA, Schmid CH, Tighiouart H, Eckfeldt JH, Feldman HI, Greene T, Kusek JW, Manzi J, Van Lente F, Zhang Y, Coresh J and Levey AS. Estimating Glomerular Filtration Rate from Creatinine and Cystatin C. *New England Journal of Medicine* 367:20-29, 2012.
8. Stevens LA, Zhang Y and Schmid CH. Evaluating the performance of equations for estimating glomerular filtration rate. *Journal of Nephrology* 21L 797-807, 2008.
9. Harrell FE, Lee KL and Mark DB. Tutorial in Biostatistics: Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Statistics in Medicine* 15: 361-387, 1996.
10. Hastie TJ. Generalized Additive Models in *Statistical Models in S*. Ed. JM Chambers and TJ Hastie, Chapman and Hall, 1993.
11. Schmid CH. Meta-Analysis, in process.

Other readings may be provided during the course.

**Schedule**

**Week 1: (Sep 4)**

**Introduction: Course overview and expections, data sets, projects, homework**

**Reading Assignment: BEG 1-2; DATA 1-2; Paper #6**

**Week 2: (Sep 9, 11)**

**Data Analysis: *Exploratory Data Analysis***

*Descriptive statistics, graphics, data sets, variable coding, ranking, ordering*

**Computing: *R Programming environment***

*Calculating environment, scripts, functions, types of data, variables, input/output, object oriented programming*

**Reading Assignment: BEG 3; DATA 4-7; Paper #6**

**Week 3: (Sep 16, 18)**

**Data Analysis: *Organizing Data***

*Designing a study; Collecting the right data; Data collection forms, Surveys, Codebooks; Documentation*

**Computing: *Basic types of data***

*Input/output, factors, subscripts, characters, dates, tables*

**Reading Assignment: DATA 8-9**

**Week 4: (Sep 25 and TBD; CS away Sep 23)**

**Data Analysis:*****Data manipulation and management, Data checking***

**Computing: *Data manipulation and reshaping data***

**Reading Assignment: : BEG 4; ANALYSIS 1-3**

**Week 5: (Sep 30, Oct 2)**

**Data Analysis: *Fitting linear models***

**Computing: *Making tables and using R model fit objects***

**Reading Assignment: BEG 5&7; GRAPH 1-3**

**Weeks 6-7: (Oct 7, Oct 16; CS away Oct 9; University holiday Oct 14)**

**Data Analysis: *Displaying model output effectively***

*Residual plots, predictive plots*

**Computing: Graphics**

*Traditional R graphics; multiple plotting figures, customization, principles of good graphics, annotation*

**Reading Assignment: GRAPH 4**

**Week 8: (Oct 21, 23)**

**Data Analysis: *Regression Diagnostics***

*Identifying influence points, outliers and leverage; model checking; identifying coding errors; prediction*

**Computing: Advanced Graphics**

*Lattice and ggplot, trellis graphics*

**Reading Assignment: ANALYSIS 4, Paper #7**

**Week 9: (Oct 28, 30)**

**Data Analysis: *Improving model fit***

*Transformations; Centering, Prediction*

**Computing: *Functions and* *Efficient coding***

*Modularization, making tables; vectorizing computations; efficient programming*

**Reading Assignment: BEG 6, Papers #8-9**

**Week 10: (Nov 4, 6)**

**Data Analysis: *Model validation***

*Performance measures (Bias, precision, accuracy), internal and external validation, cross-validation, split samples*

**Computing: *Programming structure; Writing functions***

*Looping, memory, file structures, arguments, functions*

**Reading Assignment: ANALYSIS 5, Paper #10**

**Week 11: (Nov 11,13)**

**Data Analysis: *Model Complexity***

*Nonlinear functions, interactions, splines, generalized additive models, factor coding, missing data, bias-variance tradeoff*

**Reading Assignment: ANALYSIS 7-8**

**Week 12: (Nov 18, 20)**

**Data Analysis: *: Simulation***

*Principles, random number generation, predictive models, simulating sample size*

**Computing: R random number generation**

**Reading Assignment: Efron Ch 2, 6, 7, 9, 10,12-14**

**Week 13 (Nov 25,27)**

**Data Analysis: *Bootstrap***

*Purposes, different flavors*

**Computing: R bootstrap package**

**Reading Assignment: Efron Ch 2, 6, 7, 9, 10,12-14**

**Week 14: (Dec 2,4)**

**Data Analysis: *Experimental design***

*Choosing controls,* *Randomization, Replication, Blocking/stratification, Concealing allocation, Blinding, Cross-overs, Cluster designs*

**Reading Assignment: Paper #11**

**Week 15: (Dec 9, 11)**

**Data Analysis: *Meta-analysis***

*Combining information, reconciling study protocols and coding, aggregate and individual patient data, forest plots, bias*

**Final Exam: Dec 16-18 (Take home)**