StatR 201: Data Analysis and Statistical Modeling with R

Class Master Plan and Syllabus

Instructor: Assaf P. Oron, *assaf@uw.edu*

UW Educational Outreach, Winter 2013

Lecture: Thursdays, Jan. 10 – Mar. 21, 5:30-8:30 PM PST, Lowe 216

Online Office Hours: Tuesdays, Jan. 15 – Mar. 26, 7:00-8:00 PM PST

Objectives

StatR 201 continues StatR 101 (Introduction to Statistical Analysis with R), with a focus on **statistical models and prediction methods.** Rather than theoretical rigor, the emphasis will be on conceptual understanding of the problems, of the various tools, and of their advantages and limitations. A key distinction is between *inference* and *prediction* problems. The course will place greater focus on the latter. Students will learn the fundamentals of how to approach modeling problems in a responsible and competent manner, and how to apply the acquired tools in realistic situations.

On the programming front, much of the skill development will be directly dictated by the statistical problems outlined above. We will expand upon the basic skills acquired in StatR101, closing gaps and ensuring students have a well-rounded R proficiency. Additionally, there will be a focus on data handling and manipulation, on checking code and modeling methods using simulation, on reproducible analysis using scripts, and on production-quality graphics and publishing, using advanced R packages such as `lattice, ggplot2, Sweave` and `knitr`. The latter two will be accompanied by a brief primer on equation typesetting in LaTex.

Textbooks

For R language reference, students may continue referring to the StatR101 books they have found most useful. However, at the level of StatR201 and beyond, the best resources are found online – via help pages, discussion forums and direct email interaction with other R researchers and with package maintainers. Students will learn to use the resources of the R community, with the intent of becoming contributing members.

For statistical content, there are two strongly recommended books. Both are often used as textbooks in graduate statistics programs. As such, they contain mathematical derivations at a level and quantity that goes far beyond what is needed for our course. Unfortunately, since these are advanced tools, there is no reasonable text I'm aware of, that properly presents the tools without this level of math. Therefore, when using the books we will focus mostly on the conceptual themes and discussions presented there. Students can refer to the detailed mathematics when needed, during this course or in the future. Both books will remain relevant to a good chunk of the material in StatR 301. The books are:

1. Dobson and Barnett, **An Introduction to Generalized Linear Models**, 3rd Ed., 2008.
2. Hastie, Tibshirani and Friedman, **The Elements of Statistical Learning**, 2nd Ed., 2009.

(book #2's PDF version is available online for free, from the authors, at [http://www-stat.stanford.edu/~tibs/ElemStatLearn/)](http://www-stat.stanford.edu/~tibs/ElemStatLearn/)

You might also consider a dedicated modern regression textbook, such as Montgomery et al. However, these tend to be bulky and pricy *(Montgomery et al. a bit less so; this is not an endorsement though).*

Class Format and Grading

There will be 11 Thursday lectures (Jan. 10 – Mar. 21), and 5-7 homework assignments. On Week 12, a final project will be due. There will be 11 online Tuesday office hours on Weeks 2-12. Assignment grading will follow the policies established in StatR101, namely: each homework will receive full, partial (=half) or no credit, mostly according to **evidence for acceptable and independent effort** put into solving each problem.

Successful Completion Requirements for the Course:

1. At least 75% cumulative credit from all homework assignments combined,

AND

2. Acceptable completion of the final project.

The exact requirements for the final project will be provided when the project is announced during the course.

Weekly Plan

DISCLAIMER: It the first time ever this certificate program is being taught. Also, R is a young and rapidly evolving software universe. As such – and as always – all plans are subject to change.

Week	Statistical Topics
1	**Introductions; Regression Refresher.** Class and Sequence overview. Q&A. Multiple regression: the concept, interpretation, diagnostics, sanity checking.
2	**Regression: what could possibly go wrong? (1)** Some specification do's and don'ts. Robustness/sensitivity of regression. Residual analysis and outlier triage.
3	**Maximum-Likelihood Estimation, and its application to Regression.** Estimation roles/goals. Likelihood as the Statistical Mainstream: Motivation. Formulation. Examples. Properties. Likelihood-Ratio Testing for regression. Robust and quantile regression.
4	**Generalized Linear Models (GLMs).** Motivation. Formulation. Common examples (logistic, Poisson, multinomial/ordinal). Likelihood vs. quasi-Likelihood. Nonlinear effects; spline terms in regression.
5	**Regression as a prediction tool.** Inference vs. Prediction. Gauging prediction. **Cross-validation.** Evaluating confidence-interval coverage. Inference tools for model selection: Nested tests vs. AIC and BIC. Practical Model-averaging schemes: election polls as an example.
6	**Model-Selection/Dimension-Reduction for prediction.** The problem from various perspectives. Model-selection: subset all/forward/backward. Ridge Regression and Lasso: motivation, use, examples, properties. Dimension-reduction: Principal components, PLS, SVD, projection pursuit; ensuring "honest" cross-validation.
7	**Classification (1).** The problem. Analogy to regression. Simpler methods: GLM/regression, nearest-neighbor, LDA. Tree methods. Density estimation and Naive Bayes.
8	**Classification (2).** More sophisticated methods: Dimension-reduction approaches to classification. CART, Boosting, Forests, Neural networks, SVM.
9	**R:WCPGW?(2).** Revisiting the motivation for regression: bias-variance tradeoff. Prediction-oriented curve fitting: nonlinear models, kernels, LOESS, GAM. A glimpse into correlated data and robust standard errors (?)
10	**Clustering (1).** Supervised vs. unsupervised learning. Similarities and differences with classification/regression. Basic methods: hierarchical, K-means, etc.
11	**Clustering (2).** Model-based. Attaching significance. Other advanced methods. Use of dimension reduction in clustering.

On-deck topics for Course 3:

Statistics – Bayesian concepts and methods; Markov chains. Mixed models, time series and other tools for correlated data. Bootstrap and related methods. Spatial modeling(?) High-dimensional methods for genetics and similar fields(?)

Programming – MCMC and WinBUGS, Spatial graphics, package creation, database interfacing, incorporating code from other languages. Bioconductor(?)