36-490 Undergraduate Research

Spring 2010

Vital Information

Location: Scaife Hall 212
Time: Monday and Wednesday, 10:30–11:50
Webpage: http://www.stat.cmu.edu/~cshalizi/490

Instructors:	
Brian Junker	Cosma Shalizi
132B Baker Hall	229 C Baker Hall
268-8874	268-7826

{brian,cshalizi}@ stat.cmu.edu

Office Hours: by appointment.

Course Description

In this course, students work in teams on a semester-long data analysis problem. A key objective of the course is to expose students to the variety of challenges faced by the data analyst. Students research the scientific background of their problem, consult with subject-area scientists, and communicate their methods and results both in writing and in class presentations. At the end of the semester, each team presents a poster of their project at the "Meeting of the Minds" Undergraduate Research Symposium (http://www.cmu.edu/uro/MoM/). See the attached document for a listing of projects. There will be presentations during the first two weeks describing them further.

In 490, we will extend the skills in exploring data, building and fitting models, investigating model assumptions, interpreting results from statistical models, and report writing, that were begun in 401, but will do it with real data to solve real scientific problems. We will work *concurrently* on two tracks: **weekly lectures** on statistical techniques that may be useful in your projects, and **regular project meetings** with us, within your groups, and with your faculty investigators.

Weekly Lectures Approximately once a week (most Mondays) we will provide a short lecture on a topic that we think may be useful for your projects. Usually these topics will be statistical ones (previous topics have included categorical data analysis, missing data and non-response bias, clustering, factor analysis, Markov models, etc.). Most lectures will come with short homework assignments: install a package in R, try a small data analysis on a particular data set, read a paper and discuss in the next class, etc.

While we do encourage you to discuss the homework assignments with each other, the work you hand in must be your own. You must not copy mathematical derivations, computer output and input, or written descriptions from anyone or anywhere else, without reporting the source within your work. Please review the CMU Policy on Cheating and Plagarism at
http://www.cmu.edu/policies/documents/Cheating.html.

Regular Project Meetings The projects will consume the majority of your time. Each group will get one of the instructors as a facilitator. Instead of lectures, most Thursdays you will have a group meeting with your facilitator. You should also plan on meeting at least once a week within your project group, and at least once a month with your faculty investigator.

During the semester, each group will make brief presentations to the whole class on the progress of their projects. Each group member must participate in each of these presentations. The complete project work will be presented in an end-of-the-year poster session. Here are some important dates:

Milestone	**Date**	**Notes**
Slide Presentation I	Mar 1 & Mar 3	20 min. All group members participate.
Slide Presentation II	Mar 29 & Mar 31	15 min. All group members participate.
Registration	Apr 1	"Meeting of the Minds"
Paper Draft	Apr 5	Due at 10:30AM.
Draft Poster	Apr 26	Due at 10:30AM.
Final Paper	Apr 28	Last regular course meeting.
Final Poster	May 05	"Meeting of the Minds"

Each group must turn in a formal, written report on the last day of class (Wednesday, 28 April 2010). A draft of the written report is due Wednesday, 21 March 2010. There will be no exams for this class, but several of the lectures will have associated, written homework assignments.

Two or three times during the semester, each student will be asked to assess the contribution of each group member to the team effort, and this will be factored into your project grade.

Course Mechanics

Physically Disabled and Learning Disabled Students The Office of Equal Opportunity Services provides support services for both physically disabled and learning disabled students. For individualized academic adjustment based on a documented disability, contact Equal Opportunity Services at eos@andrew.cmu.edu or (412) 268-2012.

Texts The *recommended* texts for the course are:

- W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, $4^{mathrmth}$ edition, Springer-Verlag, 2002.
- R. A. Day and G. Bastel, *How to Write and Publish a Scientific Paper*, 5^{th} edition, Greenwood Press, 2006.

This is no *required* text, but we think most of you will find these useful references to own as you do projects in 490 and beyond. We will also assign additional reading from other sources. Some other references you might find useful for this course are listed at the end of the syllabus.

Prerequisites You should have passed 36-401 (or equivalent). You should also be enrolled in 36-402. **If you do not fit into this category, please see one of us immediately to see whether it is appropriate for you to be in this course.** We expect you to have:

- a working knowledge of regression, model fitting and diagnostics as taught in 36-401
- a working understanding of the R statistical programming language
- access to a working, up-to-date installation of R. R can be run on Linux, Windows or Macintosh PC's, and it also exists on Andrew computers. It is best if you install R on your own PC since then you can easily install prewritten packages from the Web that extend R's capabilities.
- a strong desire to participate in group data analysis projects and classroom discussions.

Grading

Homework:	15%
Participation during class discussion:	10%
Participation during group project meetings:	10%
Oral Presentations:	15%
Written Report:	30%
Poster Presentation:	20%
Total:	100%

Computing, Data Sets, and Correspondence

- Homework assignments, data sets, and supplemental materials will be linked to the course website.
- Feel free to send us email with questions or comments or to schedule special appointments.

Other Useful References

Online Tutorials/References on Statistical Software

For Using R:
http://cran.r-project.org/other-docs.htm
For Using SAS:
Introduction (from UCLA):
http://www.ats.ucla.edu/stat/sas/
Regression Tutorial (from UCLA):
http://www.ats.ucla.edu/stat/sas/topics/regression.htm
Logistic Regression Tutorial (from SAS):
http://www.ats.ucla.edu/stat/sas/topics/logistic.htm

Online Certification Program (PITT)
For ensuring responsible conduct in research involving human subjects:
https://cme.hs.pitt.edu/servlet/IteachControllerServlet?actiontotake=loadmodule&moduleid=1521

Useful Books on Methodology (Some Will be on Reserve at Hunt Library)

J. M. Chambers, *Software for Data Analysis: Programming with R*, New York: Springer, 2008.
R. Christensen, *Log-Linear Models and Logistic Regression (Second Edition)*, New York: Springer, 1997.
A. C. Davison, *Statistical Models*, Cambridge, England: Cambridge University Press, 2003
J. J. Faraway, *Linear Models With R*, Boca Raton: Chapman and Hall/CRC, 2005.
—, *Extending the Linear Model with R: Generalized Linear, Mixed Effects, and Nonparametric Regression Models*, Boca Raton: Chapman and Hall/CRC, 2006.
A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis (Second Edition)*, Boca Raton: Chapman and Hall/CRC, 2004.
P. Guttorp, *Stochastic Modeling of Scientific Data*, London: Chapman and Hall, 2005.
T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edition, New York: Springer, 2009.
R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 5th edition, Upper Saddle River, NJ: Prentice Hall, 2002.
J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman, *Applied Linear Statistical Models* 4th edition, New York: WCB McGraw-Hill, 1996.
R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications: With R Examples*, 2nd edition, New York: Springer, 2006
C. F. J. Wu and M. Hamada, *Experiments: Planing, Analysis, and Parameter Design Optimization*, New York: John Wiley and Sons, 2000.

References on Statistical Consulting, Scientific Writing, and Professional Ethics

M. Alley, *The Craft of Scientific Writing*, 3rd edition, New York: Springer-Verlag, 1996. There are many examples related to the book at http://www.writing.engr.psu.edu/.
C. Chatfield, "Avoiding statistical pitfalls", *Statistical Science* **6** (1991): 249–252, http://www.jstor.org/pss/2245416
D. J. Finney, "Ethical aspects of statistical practice", *Biometrics* **47** (1991): 331–339, http://www.jstor.org/pss/2532519
G. D. Gopen and J. A. Swan, "The Science of Scientific Writing", *American Scientist* **78** (1990): 550–558, http://www.americanscientist.org/issues/num2/the-science-of-scientific-writing/1
W. G. Hunter, "The practice of statistics: The real world is an idea whose time has come", *American Statistician* **35** (1981): 72–76, http://www.jstor.org/pss/2683144
R. E. Kirk, "Statistical consulting in a university: Dealing with people and other challenges", *American Statistician* **45** (1991): 28–34, http://www.jstor.org/pss/2685235
R. Tweedie, "Consulting: Real problems, real interactions, real outcomes", *Statistical Science* **13** (1998): 1–29, http://www.jstor.org/pss/2676708
D. A. Zahn and D. J. Isenberg, "Nonstatistical aspects of statistical consulting", *American Statistician* **37** (1983): 297–302, http://www.jstor.org/pss/2682767 J. M. Williams, *Style: Toward Clarity and Grace*, Chicago: University of Chicago Press, 1990