**01:960:486 Index 12976 – COMPUTING AND GRAPHICS IN APPLIED STATISTICS**

**ARC 107, Tuesday and Friday (12:00 noon – 1:20 PM)**

**Instructor:** Jack Mardekian, PhD

Recently retired from Pfizer Inc

**Contact information:** 908.304.8400 (cell)

848.445.7647 (office)

Rutgers e-mail: [mardekia@stat.rutgers.edu](mailto:mardekia@stat.rutgers.edu)

**Office Hours:** Hill Center Rm 465: T and F 10 am – 11 am

Before and after class (I am usually outside classroom by 11:30 am)

By appointment is available, e-mail anytime

**Course Description in Catalog**:

Computing and Graphics in Applied Statistics (3)  
Prerequisite: Level II Statistics. Use of various computer-based techniques, including graphical, to understand and interpret data. Exposure to basic analysis of categorical, time-series and multivariate data in the applied areas such as biostatistics, quality control and others.

**Course Objective:**

An introduction to statistical computing and data analysis designed to provide hands on experience with statistical software. Students will work with data sets of varying size to develop the ability to generate analyses, interpret output and summarize results both written and verbally. Students will be introduced to R programming and work with a variety of available R procedures.

**Important Considerations:** You will be writing and executing R programs for this class. R programs will be provided that you can use and modify. You learn by writing and fixing your R code. You will need to be able to refer to R manuals and examples that are available through the Internet to successfully complete homework assignments. We will also be reviewing SAS output but you will not need to execute SAS (SAS is no longer listed as a standard offering by the Computer Labs).

Rutgers Office of Information Technology website: <https://oit-nb.rutgers.edu/>

**Canvas Class Site:** <https://rutgers.instructure.com/courses/39556>

Class notes, assignments, announcements will be posted on this site.

**Text Book:** *Linear Models with R*, Julian J. Faraway, Taylor, second edition, 2014

ISBN: 9781439887332

**Textbook site:** <http://people.bath.ac.uk/jjf23/LMR/>

**Rutgers Honor Pledge is in effect at all times:** “On my honor, I have neither received nor given any **unauthorized** assistance on this examination.”

**Grading (weights are tentative):**

2 In-class tests, possible quizzes (40%) (To be scheduled soon)

1 In-class Final Exam (40%) (Tentatively scheduled for Thursday, May 7, 12 noon - 3 pm)

HW assignments submitted electronically using Canvas (10%)

Class project (10%)

**Other relevant dates:**

|  |  |
| --- | --- |
| Spring Recess—University Offices Open—No Classes | Saturday, March 14 - Sunday, March 22 |
| Regular Classes End | Monday, May 4 |
| Final Exam Period | Thursday, May 7 - Wednesday, May 13 |

**Attendance and timely completion of HW assignments in Canvas are minimum requirements for passing this course.**

**Other reference material:**

The Elements of Statistical Learning. Data Mining, Inference, and Prediction, 2nd Edition, February 2009, Trevor Hastie, Robert Tibshirani, Jerome Friedman

Free book PDF available at: <http://www-stat.stanford.edu/~tibs/ElemStatLearn/>

An Introduction to Statistical Learning, with applications in R (Springer, 2013)

G. James, D. Witten, T. Hastie and R. Tibshirani

Free book PDF available at: <http://www-bcf.usc.edu/~gareth/ISL/>

**Other Useful Information:**

Example data sets can be found at: <http://archive.ics.uci.edu/ml/>

The R Project for Statistical Computing: <http://www.r-project.org/>

Rutgers computing site: <https://oit-nb.rutgers.edu/>

Display the remote desktop through your Web browser <https://apps.rutgers.edu/novnc/>

**For R material on the Web** – See [www.rseek.org](http://www.rseek.org)

also Google “An Introduction to R”

All web sites below were successfully accessed on January 16, 2020

**Resources to help you learn and use R**

<http://www.ats.ucla.edu/stat/r/>

**University of California Irvine (UCI) Machine Learning Repository** (data sets)

<http://archive.ics.uci.edu/ml/>

**Useful Statistics web sites:**

<http://www.claviusweb.net/statistics.shtml>

<http://stattrek.com/>

R passes SAS in scholarly use – June/2016

<http://r4stats.com/2016/06/08/r-passes-sas-in-scholarly-use-finally/>

Also see updated articles by the author: <http://r4stats.com/author/muenchen-bobgmail-com/>

**Career Planning?** For Today’s Graduate, Just One Word: Statistics"

<http://www.nytimes.com/2009/08/06/technology/06stats.html?_r=1&em>

**Suggestion:** Join Machine Learning and Data Science Central groups in LinkedIn

[https://www.**linkedin**.com/company/**data**-**science**-**central**](https://www.linkedin.com/company/data-science-central)

**FYI:** Selected questions from “66 job interview questions for data scientists” posted by Vincent Granville on February 13, 2013. These questions were selected as motivation for the Project Assignment. According to Vincent Granville: “These questions are mostly open-ended questions, to assess the technical horizontal knowledge of a senior candidate for a rather high-level position, e.g. Director.”

1. What is the biggest data set that you processed, and how did you process it, what were the results?

2. Tell me two success stories about your analytic or computer science projects? How was lift (or success) measured?

3. How would you come up with a solution to identify plagiarism?

4. What is better: good data or good models? And how do you define "good"? Is there a universal good model? Are there any models that are definitely not so good?

5. What is probabilistic merging (AKA fuzzy merging)?

6. How do you handle missing data? What imputation techniques do you recommend?

7. Tell me 3 things positive and 3 things negative about your favorite statistical software.

8. You are about to send one million email (marketing campaign). How do you optimize response?

9. What are the drawbacks of general linear model? Are you familiar with alternatives (Lasso, ridge regression, boosted trees)?

10. Give examples of data that does not have a Gaussian distribution, nor log-normal.

11. Why is mean square error a bad measure of model performance? What would you suggest instead?

12. What is sensitivity analysis? Is it better to have low sensitivity (that is, great robustness) and low predictive power, or the other way around? How to perform good cross-validation? What do you think about the idea of injecting noise in your data set to test the sensitivity of your models?

13. Do you know / used data reduction techniques other than PCA? What do you think of step-wise regression? What kind of step-wise techniques are you familiar with? When is full data better than reduced data or sample?

14. Are you familiar either with extreme value theory, monte carlo simulations or mathematical statistics (or anything else) to correctly estimate the chance of a very rare event?

15. How would you define and measure the predictive power of a metric?

16. What is an exact test? How and when can simulations help us when we do not use an exact test?

17. You design a robust non-parametric statistic (metric) to replace correlation or R square, that (1) is independent of sample size, (2) always between -1 and +1, and (3) based on rank statistics. How do you normalize for sample size?

18. What/when is the latest data mining book / article you read? What/when is the latest data mining conference / webinar / class / workshop / training you attended? What/when is the most recent programming skill that you acquired?

|  |  |  |
| --- | --- | --- |
| **Topics to be covered from text** |  |  |
| Linear Models with R |  |  |
|  | 6 Diagnostics | 12 Insurance Redlining — A Complete Example |
| Table of Contents | 6.1 Checking Error Assumptions | 12.1 Ecological Correlation |
|  | 6.1.1 Constant Variance | 12.2 Initial Data Analysis |
| Preface | 6.1.2 Normality | 12.3 Full Model and Diagnostics |
| 1 Introduction | 6.1.3 Correlated Errors | 12.4 Sensitivity Analysis |
| 1.1 Before You Start | 6.2 Finding Unusual Observations | 12.5 Discussion |
| 1.2 Initial Data Analysis | 6.2.1 Leverage | 13 Missing Data |
| 1.3 When to Use Linear Modeling | 6.2.2 Outliers | 13.1 Types of Missing Data |
| 1.4 History | 6.2.3 Influential Observations | 13.2 Deletion |
| 2 Estimation | 6.3 Checking the Structure of the | 13.3 Single Imputation |
| 2.1 Linear Model | Model | 13.4 Multiple Imputation |
| 2.2 Matrix Representation | 6.4 Discussion | 14 Categorical Predictors |
| 2.3 Estimating β | 7 Problems with the Predictors | 14.1 A Two-Level Factor |
| 2.4 Least Squares Estimation | 7.1 Errors in the Predictors | 14.2 Factors and Quantitative Predictors |
| 2.5 Examples of Calculating βˆ | 7.2 Changes of Scale | 14.3 Interpretation with Interaction Terms |
| 2.6 Example | 7.3 Collinearity | 14.4 Factors With More Than Two Levels |
| 2.7 QR Decomposition | 8 Problems with the Error | 14.5 Alternative Codings of Qualitative |
| 2.8 Gauss–Markov Theorem | 8.1 Generalized Least Squares | Predictors |
| 2.9 Goodness of Fit | 8.2 Weighted Least Squares | 15 One Factor Models |
| 2.10 Identifiability | 8.3 Testing for Lack of Fit | 15.1 The Model |
| 2.11 Orthogonality | 8.4 Robust Regression | 15.2 An Example |
| 3 Inference | 8.4.1 M-Estimation | 15.3 Diagnostics |
| 3.1 Hypothesis Tests to Compare | 8.4.2 Least Trimmed Squares | 15.4 Pairwise Comparisons |
| Models | 9 Transformation | 15.5 False Discovery Rate |
| 3.2 Testing Examples | 9.1 Transforming the Response | 16 Models with Several Factors |
| 3.3 Permutation Tests | 9.2 Transforming the Predictors | 16.1 Two Factors with No Replication |
| 3.4 Sampling | 9.3 Broken Stick Regression | 16.2 Two Factors with Replication |
| 3.5 Confidence Intervals for β | 9.4 Polynomials | 16.3 Two Factors with an Interaction |
| 3.6 Bootstrap Confidence Intervals | 9.5 Splines | 16.4 Larger Factorial Experiments |
| 4 Prediction | 9.6 Additive Models | 17.1 Randomized Block Design |
| 4.1 Confidence Intervals for | 9.7 More Complex Models | 17.2 Latin Squares |
| Predictions | 10 Model Selection | 17.3 Balanced Incomplete Block Design |
| 4.2 Predicting Body Fat | 10.1 Hierarchical Models |  |
| 4.3 Autoregression | 10.2 Testing-Based Procedures |  |
| 4.4 What Can Go Wrong with | 10.3 Criterion-Based Procedures |  |
| Predictions? | 10.4 Summary |  |
| 5 Explanation | 11 Shrinkage Methods |  |
| 5.1 Simple Meaning | 11.1 Principal Components |  |
| 5.2 Causality | 11.2 Partial Least Squares |  |
| 5.3 Designed Experiments | 11.3 Ridge Regression |  |
| 5.4 Observational Data | 11.4 Lasso |  |
| 5.5 Matching |  |  |
| 5.6 Covariate Adjustment |  |  |
| 5.7 Qualitative Support for Causation |  |  |

**Other course topics:**

Exploratory Data Analysis - summary statistics, graphical summaries: Stem and leaf displays, box plots, checking the shape of distributions with Quantile-Quantile plots, Monte Carlo simulation, Poissonness Plot

Smoothing techniques, Newton-Raphson technique. Maximum likelihood. Monte Carlo simulation, Bayesian analysis, Data Mining**:** Random Forests, k-fold Cross Validation

**The Comprehensive R Archive Network (CRAN)**

**What are R and CRAN?**

R is ‘GNU S’, a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc. Please consult the [R project homepage](https://www.r-project.org/) for further information.

CRAN is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R. Please use the CRAN [mirror](https://cran.r-project.org/mirrors.html) nearest to you to minimize network load.

|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | | Download and Install R Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:   * [Download R for Linux](https://cran.r-project.org/bin/linux/) * [Download R for (Mac) OS X](https://cran.r-project.org/bin/macosx/) * [Download R for Windows](https://cran.r-project.org/bin/windows/)   R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above. | | Source Code for all Platforms Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!   * The latest release (2019-12-12, Dark and Stormy Night) [R-3.6.2.tar.gz](https://cran.r-project.org/src/base/R-3/R-3.6.2.tar.gz), read [what's new](https://cran.r-project.org/doc/manuals/r-release/NEWS.html) in the latest version. * Sources of [R alpha and beta releases](https://cran.r-project.org/src/base-prerelease/) (daily snapshots, created only in time periods before a planned release). * Daily snapshots of current patched and development versions are [available here](https://stat.ethz.ch/R/daily). Please read about [new features and bug fixes](https://cran.r-project.org/doc/manuals/r-devel/NEWS.html) before filing corresponding feature requests or bug reports. * Source code of older versions of R is [available here](https://cran.r-project.org/src/base/). * Contributed extension [packages](https://cran.r-project.org/web/packages/index.html) | | Questions About R  * If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](https://cran.r-project.org/faqs.html) before you send an email. | |

**NOTE:** Rutgers Computer Labs software includes R.3.51 (MAC) and 3.5.2 (PC) and Rstudio 1.1.463