DEPARTMENT OF MATHEMATICS AND STATISTICS

**STA 580/780 R and Introductory Data Mining CRN 16177/16178 3 credit hours**

**MW 6:00-7:15 PM Wallace 444**

**Fall 2016** **Syllabus**

|  |  |  |
| --- | --- | --- |
| **Instructor Information** | **Name: Benjamin Nutter** **Office:**  **Phone:** 859-625-8398 **e-mail**: benjamin.nutter@eku.edu  MW: 5:00 - 6:00 PM Other hours by appointment. | |
| **Prerequisite** | STA 580 | STA 320, MAT 124, and CSC 160 or 174 or 177 or 190 |
| STA 780 | STA 700, CSC 730, and any Calculus class, or departmental approval |
| **Course Description** | Data set manipulation, application of statistical techniques in R, statistical programming, and data mining skills. | |
| **Required Materials** | 1. **TEXT : a. Required: R For Data Science** (Grolemund and Wickham)   <http://r4ds.had.co.nz/>  **b**. **Recommended: An Introduction to R**. (Venables and Smith)  <https://cran.r-project.org/doc/manuals/R-intro.pdf>  **c. The R Inferno (Burns)**  <http://www.burns-stat.com/pages/Tutor/R_inferno.pdf>  **d. Internet access**  **e. R (>= 3.3.0)**  **f. R Studio (>= 0.99.903)**  **g. pandoc (>= 1.17.2)** | |
| **Purpose of the Course** | Computing with data is an essential part of statistical research and practice. The R language provides a rich computing environment for working with data. R not only offers an abundance of packages that allow users the access to well-established statistical techniques, but also provides great flexibility for users to create their own methods to handle the data. Backed by the detailed documentation and support from both inside and outside of the statistical community, R has become an indispensable tool for statistical computing and therefore an integral part of a statistics program.  This course will help students develop the basic skills of conducting statistical analysis using R. A student from this course is expected to be comfortable with the R environment, be proficient to conduct data set manipulation and to apply various statistical techniques. A student is also expected to write simple programs to implement his/her idea of building statistical models. | |
| **Student Learning Outcomes** | Students who successfully complete STA 580 or STA 780 will demonstrate the ability to use R to:   1. Perform basic programming. 2. Make use of literate programming. 3. Conduct operations on numbers, vectors, arrays and matrices. 4. Apply basic statistical models and interpret the output. 5. Write simple functions to implement statistical methods. 6. Utilize basic skills of statistical data mining.   Students who successfully complete STA 780 will also demonstrate the ability to:   1. Analyze data using advanced statistical models. 2. Implement intermediate data mining skills. 3. Develop efficient computational algorithms. 4. Explore and adopt R packages for statistical analysis in different areas. | |
| **Tentative Course Outline** | **List of Topics**   1. **Introduction to R (1 week)**    1. Put R to work(download and installation; start and end an R session)    2. Get help and find R documentation    3. R environment    4. R Markdown 2. **Basic R Operations (2 weeks)**    1. Basic computation    2. Work with vectors    3. Work with arrays and matrices 3. **Grouping, Loops and Conditional Operations with R (1 week)**    1. Conditional operations: “if” statements    2. Loops and repetitive operations: “for” loops, “repeat” and “while” statements 4. **Write Your Own Function with R / Using \*apply (2 weeks)**    1. Writing basic functions    2. Frequently used functions in statistics    3. Tips for writing your own functions    4. Repeating operations with \*apply 5. **Data Management (2 weeks)**    * + 1. Basic transformations        2. Character strings        3. Exploratory Data Analysis        4. Graphics with R 6. **Statistical Analysis with R (6 weeks)**    1. Implementing basic statistical methods in R       * 1. Confidence interval and hypothesis testing         2. Linear regression         3. Analysis of variance (ANOVA)         4. Logistic regression         5. Survival analysis         6. Random forest    2. Model evaluation   a. Model diagnostics  b. Information criteria  c. Likelihood ratio tests  d. Cross validation  e. Bootstrapping   1. **Simulation and optimization (1 week)**    1. Random number generation    2. Sample size estimation    3. Sample size simulation 2. **Writing R Packages (1 week)**    1. Dependencies    2. Documentation    3. Methods | |
| **Inclement Weather Plan** | On inclement weather days, this class will meet as scheduled unless cancellations are announced. See the inclement weather plan at http://www.eku.edu/weather/richmond-campus-inclement-weather-plan | |
| **Note** | It is expected that each STA 580 student will spend on the average a minimum of **three hours** outside study for each hour of class time. It is expected that each STA 780 student will spend on the average a minimum of **four hours** outside study for each hour of class time. | |
| **Attendance Policy** | Regular class attendance is essential. Unexcused absences for more than 10% of the regularly scheduled class meetings may result in a lower course grade. The instructor may excuse an absence only when the student presents an adequate and/or documented reason within a reasonable amount of time. Such reasons usually include circumstances beyond the student's control, such as personal illness, critical illness or death in the immediate family, or participation in university-sponsored activities. Under extraordinary circumstances, this policy may be waived for individuals at the discretion of the instructor. Students will be held responsible for announcements made in class. | |
| **Student Progress** | Mid-term grades should be viewable online on **Monday, October 10, 2016**. | |
| **Add/Drop**  **Withdrawal** | Sunday, August 28, 2016, is the last day to drop this class with no tuition charge and no record of the course on the transcript. The last day to withdraw from this course with no withdrawal fee is Sunday, September 18, 2016. After that, students withdrawing from the class may incur a $150 fee. The last day to withdraw from this class is Friday, November 13, 201. For additional information about dates for partial refunds, consult the Colonel's Compass, [http://colonelscompass.eku.edu/fall-2016-deadlines-adddrop-refunds](http://colonelscompass.eku.edu/spring-2016-deadlines-adddrop-refunds%22http://colonelscompass.eku.edu/fall-2014-deadlines-adddrop-refund) | |
| **Make-Up Policy** | Student anticipating to miss an in-class quiz or exam should notify the instructor as early as possible about the incidence. A student missing an in-class quiz or exam due to unexpected reasons must contact the instructor within 1 day of the scheduled assignment to obtain permission for a make-up. In both situations, a valid reason for the absence as well as documentation is required. Lack of valid reason or documentation may cause a deduction of the student’s grade from the assignment or void the student’s right to a make-up. | |
| **Services for Individuals with Disabilities** | The University strives to make all learning experiences as accessible as possible. If you are registered with the EKU Center for Student Accessibility (CSA), please obtain your accommodation letters from the CSA, present them to the course instructor, and discuss the accommodations needed. If you believe you need an accommodation and are not registered with the CSA, please contact the office in 361 Whitlock Building by email at [disserv@eku.edu](mailto:disserv@eku.edu) or by telephone at [(859) 622-2933](tel:%28859%29%20622-2933). Upon individual request, this syllabus can be made available in an alternative format.  A student with a “disability” may be an individual with a physical or psychological impairment that substantially limits one or more major life activities, to include, but not limited to: seeing, hearing, communicating, interacting with others, learning, thinking, concentrating, sitting, standing, lifting, performing manual tasks, working. Additionally, pregnancy accompanied by a medical condition(s), which causes a similar substantial limitation, may also be considered under the Americans with Disabilities Act Amendments Act (ADAAA). | |
| **Department Policies** | * During class, cell phones and pagers must be turned off or set to a silent mode. * During resource-limited activities, such as in-class exams, students may not use the calculator function of a wireless communication device such as a cell phone or PDA. * Grades are not given out over the phone. They may be posted on Blackboard. * Any student enrolling in a multiple section course for which s/he has already received a grade of “D”, “F”, or “W” from the instructor who is teaching the section may change to a section taught by a different instructor by seeing the Chair of the Department of Mathematics and Statistics. This change must be completed by the end of the drop/add period. | |
| **Course Requirements and Grading Policy** | **Scoring**: Homework 25%  In-class Exams 5%  Midterm Exam/projects 2@20% each  Final Exam/project 30%  **Scoring Details:**  **Homework:** Homework will be assigned every one or two weeks.All the homework assignments are strictly individual and will be given in the assignment section of Blackboard.  Midterm Exam/projects: The projects will be take-home and can be either individual or group. They will be announced during the lecture. The instructions for the projects will be posted in the assignment section of Blackboard.  **Final Exam/project**: The final exam/project will be announced at least two weeks before the end of classes and will be completed in the week of final exams.  **Grading Scale:** The grading scale for the course is   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | % | < 60 | 60-69 | 70-79 | 80-89 | 90-100 | | Grade | F | D | C | B | A | | |
| **Additional Requirements For STA 780 Students** | STA 780 students will be held to a higher standard than those enrolled in STA 580. They will be expected to develop stronger programming skills and to implement more advanced statistical techniques in R. They will also be required to explore more topics in data mining than STA 580 students.  Extra problems will be assigned to students enrolled for STA 780 on some of the homework assignments and projects. On exams, there will be different problems for STA 580 and STA 780 students.  STA 780 students will also be required to complete an extra project that requires them to explore and implement one topic in data mining that is related to the course materials but not covered in lectures. The extra project will be due in the last week of class. | |
| **Academic Honesty and Responsibility** | 1. Anyone violating the usual standards for academic honesty, for example, anyone attempting to obtain or exchange information regarding any quiz or test, or anyone using a fraudulent excuse to qualify for a make-up, may receive a course grade of “F”. Cheating includes buying, selling, or otherwise fraudulently obtaining copies of examinations or assignments for the purpose of improving one’s academic standing. During examinations it includes receiving information from other students or other students’ exams and referring to unauthorized notes or other written information in any form including electronic (e.g., information stored in graphing calculators). 2. Copying or working together **is not** acceptable for computer assignments. If you are caught copying or working together, you will definitely get a zero for the assignment, and you may get an “F” for the course. This includes the person who allows another person to copy his or her assignment as well as the person who copied, or all students working together. Working together includes, but is not limited to, discussing wording of an answer, discussing any computations or numerical answers, etc. If you have **any** questions regarding a computer assignment, you should ask your instructor. **Do not ask other students for help.** 3. Anyone behaving in a disruptive manner or refusing to follow the usual standards for academic behavior may be barred from attending class and may receive a course grade of “F”. 4. Students are advised that EKU’s Academic Integrity policy will be strictly enforced in this course. The Academic Integrity policy is available at <http://studentrights.eku.edu/academic-integrity-policy>. Questions regarding the policy may be directed to the Office of Academic Integrity. | |
| **Official E-mail** | An official EKU e-mail address is established for each registered student, each faculty member, and each staff member. All university communications sent via e-mail will be sent to this EKU e-mail address. | |
| **Selected Resources for R** | * *Advanced R* by Hadley Wickham (<http://adv-r.had.co.nz/>) * *R packages* by Hadley Wickham (<http://r-pkgs.had.co.nz/>) * *The R Graphics Cookbook* by Winston Chang (O'Reilly, 2013; ISBN 978-1449316952) * [*Software for Data Analysis: Programming with R*](http://www2.lib.purdue.edu:2048/login?url=http://dx.doi.org/10.1007/978-0-387-75936-4) by John M. Chambers (Springer, 2008) * *Introductory Statistics with R*, 2nd Edition by Peter Dalgaard (Springer, 2008) * *Data Analysis and Graphics Using R*, 2nd Edition by John Maindonald and John Braun (Cambridge, 2003) * *Using R for Introductory Statistics* by John Verzani (Chapman & Hall/CRC, 2005) * *The R Book* by Michael J. Crawley (Wiley, 2007) * *Statistics: An Introduction using R* by Michael J. Crawley (Wiley, 2005) * *A Handbook of Statistical Analyses Using R* by Brian Everitt and Torsten Hothorn (Chapman & Hall/CRC, 2006) * *Statistical Computing with R* by Maria L. Rizzo (Chapman & Hall/CRC, 2008) * *A First Course in Statistical Programming with R* by W. John Braun and Duncan J. Murdoch (Cambridge, 2007) * [*Data Manipulation with R*](http://www2.lib.purdue.edu:2048/login?url=http://dx.doi.org/10.1007/978-0-387-74731-6) by Phil Spector (Springer, 2008) * *Introduction to Probability with R* by Kenneth Baclawski (Chapman & Hall/CRC, 2008) * *Probability and Statistics with R* by María Dolores Ugarte, Ana F. Militino, and Alan T. Arnholt (Chapman & Hall/CRC, 2008) * [*R: A language for data analysis and graphics*](http://www2.lib.purdue.edu:2111/stable/1390807), by Ross Ihaka and Robert Gentleman, Journal of Computational and Graphical Statistics, 5(3):299-314, 1996 * [*Data Analysts Captivated by R's Power*](http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html), by Ashlee Vance, New York Times, January 6, 2009 * [*R for SAS and SPSS Users*](http://rforsasandspssusers.com/), by Robert A. Muenchen (Springer, 2008) | |