Regression and Multivariate Data Analysis

STAT-GB.2301 / STAT-UB.0017
Jeffrey S. Simonoff
Office: KMC 8-54 (Office hours: Wednesdays 2:00-4:00 PM and by appointment)
Phone: (212) 998-0452
FAX: (212) 995-4003
e-mail: jsimonof@stern.nyu.edu
WWW: http://people.stern.nyu.edu/jsimonof/classes/2301

IMPORTANT NOTE: This web page refers to the Regression and Multivariate Data Analysis class to be taught during the Spring 2016 semester.

This is a data-driven, applied statistics course focusing on the analysis of data using regression models. It emphasizes applications to the analysis of business and other data and makes extensive use of computer statistical packages. Topics include simple and multiple linear regression, residual analysis and other regression diagnostics, multicollinearity and model selection, autoregression, heteroscedasticity, regression models using categorical predictors, and logistic regression. All topics are illustrated on real data sets obtained from financial markets, market research studies, and other scientific inquiries. The goal of the class is that students begin to develop the skills to be able to collect, organize, analyze, and interpret regression data.

If you are a non-Stern NYU student, there are certain procedures that you must follow in order to register for the course. Please click here for details if you are a graduate student, and click here for details if you are an undergraduate student.

Texts

Samprit Chatterjee and Jeffrey S. Simonoff, *Handbook of Regression Analysis*, John Wiley and Sons (2013). [Highly recommended, but not required; you can do all of the work required for the class without it. In any event, I believe that it is a useful applied guide to have.]

The course grade will be based on homeworks/projects **only**. Grades will be determined based on a class-wide curve (that is, there will **not** be separate curves for undergraduates and graduate students). The course will be **very** heavily computer oriented; if you have not used a statistical package before, you may be in for some rough going. The "official" package for the course is Minitab, which is available on the Stern network, for purchase at the bookstore, and for rental through the website http://www.onthehub.com (I **highly** recommend that you either purchase or rent the package). You may use any package you wish, on any machine that you wish, **as long as it performs the necessary calculations**; any deficiencies on the part of the package **are the responsibility of the student**. I can provide additional support for Minitab and R, but relatively little for SAS, and none at all for SPSS, Stata, Systat, and STATISTICA (although these packages are able to perform all of the necessary modeling methods for this class). Excel will **not** be an acceptable tool for analyses in this class, and the student version of Minitab is missing some necessary techniques that are included in the full version. I have put S-PLUS/R and SAS code for the different handouts up on the class web site; for S-PLUS/R click here, and for SAS click here.

As you know, the Stern schedule for night (Langone) classes is noticeably different from that for day classes. In particular, classes only meet for 12 weeks, rather than for 13 weeks plus a final/extra class session. **Please note that the first day of class is Wednesday, February 10, 2016.** It is **crucially important** that all students review basic regression material **before** the first class. Please see the material under *Required work before first class session* below.

Some of you might not have very much experience in reading or writing statistical reports, a skill that you will need for this class. The Chatterjee, Handcock and Simonoff (CHS) Casebook gives many examples of such reports. I urge you to read some of the cases that appear in the book to see what such reports look like if you have concerns about this. Examples include the first two cases in each of the sections **Data analysis**, **Applied probability**, **Statistical inference**, and **Regression analysis**. You will find that my reports are somewhat "chatty" - it's perfectly appropriate (even desirable) for you to write such reports for this class, but you should be aware that the reports you might write for other classes might need to be more factual and to-the-point. An excellent way to get "up to speed" in your statistical computing is to work through these cases on the computer. If you are comfortable with your ability to write such reports, you will probably find the Casebook to be of little use to you.

For most assignments, you will be responsible for obtaining your own data. Do **not** merely take data from a textbook; obtain your data from original data sources. You will be required to provide complete source information for your data (a URL if the data come from the World Wide Web, or a photocopy of the appropriate page(s) if the data come from a printed source). Generally speaking, you will have roughly two weeks to complete each assignment. Assignments **must** be typed or word processed; handwritten assignments will **not** be accepted.

The Stern Code of Conduct states that you commit to "Exercise integrity in all aspects of our academic work including, but not limited to, the preparation and completion of exams, papers and all other course requirements by not engaging in any method or means that provides an unfair advantage." Further, you commit to "Refrain from behaving in ways that knowingly support, assist, or in any way attempt to enable another person to engage in any violation of the Code of Conduct. Our support also includes reporting any observed violations of this Code of Conduct or other School and University policies that are deemed to have an adverse effect on the NYU Stern community." This applies to this class in the following specific ways (in addition to general prohibitions on cheating, plagiarism, and so on):

- All data analyses must be done independently. I will be happy to answer questions about your analyses (either in person or via e-mail), but you'll probably find that as the semester goes on I'll be increasingly likely to answer "What do **you** think?" to many questions! Please do not give me preliminary drafts of your homework to check. I will answer specific questions (if possible), but will not review drafts to provide general comments or reactions. You can certainly get help from classmates or people outside of the class on computational issues (how to do something in Minitab, for example), but not on conceptual and/or substantive statistical issues. In particular, it is **not** permitted for you to show your assignment to anyone else in the class, or for you to look at anyone else's assignment, whether that assignment is from this semester's class or a previous class.
- Data sets cannot be taken from a source where a similar analysis is already given.

Violation of these conditions can lead to loss of all credit for the assignment involved at a minimum, with more severe sanctions possible after consultation with the Dean's Office.

A friendly piece of advice: **don't hand in the assignments late!** That is the quickest way to get in trouble in a course like this. An assignment is considered late if it is turned in after I have left Stern for the day on the day that it is due. There will be progressively larger penalties for increasing amount of lateness of an assignment (2 points out of 10 up to one week late, 4 points out of 10 up to two weeks late). **No** assignments will be accepted for credit more than two weeks late. Work responsibilities in general, including work-related travel in particular, will **not** be accepted as an excuse for lateness of an assignment; it is your responsibility to get the assignment to me on time, even if you are not at Stern that day. For some of the questions and/or assignments, I may give out an answer sheet when I return graded assignments; late assignments will **not** be accepted for credit after answers have been distributed. **Don't** wait until the last minute to do an assignment, as you might find that access to School (or any other)

computing facilities is difficult or impossible (the network might be down, your laptop's hard drive might crash, or your printer might run out of ink); such lack of access will **not** be accepted as an excuse for lateness.

I have had complaints from students in the past regarding distractions caused by students using laptops in class. If you want to use a laptop in class for note taking, or to follow along with the discussion or statistical analyses done in class, I ask that you sit in the back of the classroom. Of course, surfing the web, answering e-mails, instant messaging, etc., are not appropriate uses of a laptop (or any electronic device) under any circumstances.

If you have a qualified disability and will require academic accommodation during this course, please contact the Moses Center for Students with Disabilities (CSD, 998-4980) and provide me with a letter from them verifying your registration and outlining the accommodations they recommend.

The final grade for the course will be based on the grades on the assigned homeworks **only**; there will be no opportunities for makeup or extra credit work, and an incomplete grade for the course will **not** be considered simply to make up assignments that were not done. Thus, assignments for which you receive no credit will have a strong detrimental effect on your grade, and as few as two such assignments could result in a failing grade in the course. The actual curve used in the course will depend on the performance of the class, but in the past the cutoff for A grades (A and A-) has been roughly 8.5 (out of 10), while the cutoff for B grades (B+, B, and B-) has been roughly 7.5 (there is no guarantee that these cutoffs will apply this semester, however).

Most importantly - **THIS COURSE IS LIKELY TO BE TIME-CONSUMING!** If you're taking a particularly heavy course load this semester, or are going to be doing a lot of traveling (work-related, for example), this is probably not the course for you! In particular, since it takes time to build up the knowledge necessary for adequate multiple regression analysis, the homeworks will be relatively widely separated in the first half of the semester, but will come more rapidly in the second half.

I will be giving out handouts in many of the class sessions during the semester. If you know that you're going to miss class, you can get in touch with me beforehand, and I will save copies for you. I cannot guarantee to have copies left over for you after class is over - you will probably have to get copies from a classmate. You should make every effort not to miss classes, however, since the material covered in class will be far more relevant to you than is material in the textbook.

Prerequisite: Introductory statistics core course. More generally, the prerequisite is an introductory statistics class that includes discussion of descriptive statistics and univariate statistical inference (confidence intervals, prediction intervals, and hypothesis testing), and exposure to simple regression methods.

Required work before first class session: I will assume a basic understanding of the simple regression model from the beginning of the class. You should review this material from your introductory statistics course **before** the first class session. You should download, print out, and read the following handouts: Regression - the basics and Purchasing power parity - is it true?. You are responsible for all of the material in those handouts, although we will **briefly** discuss them in class. You should also download Homework 1 and answer all of the questions. I will give out the answers to these questions on the first day of class.

Syllabus

Chapters refer to the Chatterjee and Simonoff book. Corresponding class sessions given are only

approximate.

Classes 1-3

1. Simple and multiple regression - Chapter 1

Classes 3-4

2. Checking assumptions of regression - Chapters 1, 2, 3

Classes 4-8

3. Addressing violation of assumptions: choosing the correct predictors (model selection), autocorrelation - Chapters 2, 4, 5

Classes 8-9

4. Analysis of variance and covariance and nonconstant variance - Chapters 6, 7

Classes 10-12

5. Modeling group membership: logistic regression - Chapters 8, 9