

Regression and Multivariate Data Analysis

STAT-GB.2301 (B90.2301) STAT-UB.0017 (C22.0017)

Jeffrey S. Simonoff

Office: KMC 8-54

Phone: (212) 998-0452

FAX: (212) 995-4003

e-mail: jsimonof@stern.nyu.edu

Class meeting time: Tuesdays/Thursdays, 1:30 – 2:50 PM, KMC 3-70

WWW: <http://www.stern.nyu.edu/~jsimonof/classes/2301>

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.

Texts:

Samprit Chatterjee and Ali S. Hadi, *Regression Analysis By Example, 4th. edition*, John Wiley and Sons (2006). [Highly recommended, but not required; you can do all of the work required for the class without it. In any event, it is an excellent general reference to have.]

Samprit Chatterjee, Mark S. Handcock and Jeffrey S. Simonoff, *A Casebook for a First Course in Statistics and Data Analysis*, John Wiley and Sons (1995). [Optional; see discussion below.]

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 online, for rent, and for purchase at the bookstore (I **highly** recommend that you either rent or purchase 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**. Note that the “student version” of **Minitab cannot** do all of the analyses required for the class. I can provide additional support for **Minitab**, **S-Plus**, 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.

Class will meet according to the Stern graduate calendar, **not** the standard university calendar. See the Stern website for details on the differences between the two calendars. 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 **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 few 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 from when it is given out, although in some cases it will be expected that material that is covered in class after the assignment is handed out (but before it is due) will be used by you in the 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):

1. 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 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.

2. 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 bigger penalties for increasing amount of lateness of an assignment (1.5 points out of 10 up to one class late, 3 points up to two classes late, 6 points up to three classes late). **No** assignments will be accepted for credit more than three classes late under **any** circumstances. 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. **Please** keep in mind that you will **not** be graded based on how "exciting" your data are, but rather on the quality of your analysis. **Don't** waste time trying to find the "perfect" data set; you're working on a homework assignment, not a master's thesis. If you find that you're spending more time finding data than you are on analyzing it and writing up the analysis, you are allocating your time incorrectly.

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.

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.

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 lower cutoff for A grades (A and A-) has been roughly 8.5 (out of 10), while the lower 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 Hadi book. Corresponding class sessions given are only approximate.

Classes 1-4

1. Review of basic regression concepts — Chapters 1, 2

Class 4

2. Matrix approach to regression — Appendix to Chapter 3

- Classes 4–7*
3. Multiple regression — Chapter 3
- Classes 8–10*
4. Checking assumptions of regression — Chapter 4
- Classes 11–16*
5. Addressing violation of assumptions: choosing the correct predictors (model selection), autocorrelation — Chapters 6, 8, 9, 11
- Classes 17–21*
6. Analysis of variance and covariance and nonconstant variance — Chapters 5, 7
- Classes 22–27*
7. Modeling group membership: logistic regression — Chapter 12

The following is a list of the handouts that will be given out in class, separated by broad coverage.

Simple regression

- Regression — the basics
- Purchasing power parity — is it true?

Multiple regression, including use of partial F -tests

- Multiple regression
- Getting what you pay for: dinner prices in Manhattan
- Mortgage rates
- Purchasing power parity, revisited

Regression diagnostics

- Regression diagnostics

Transformations

- Transformations in regression
- Predicting total movie grosses after one week
- Modeling Lowe's sales

Model selection

- Estimating a demand function

Time series data

- Ordinary least squares estimation and time series data
- Estimating a demand function — it's about time
- Eruptions of the Old Faithful Geyser

Analysis of variance and covariance

One-way ANOVA

Assessing the credit risk of fixed income securities

Two-way ANOVA

Modeling television viewership

Analysis of covariance

CAPM: Do you want fries with that?

Logistic regression

Logistic regression — modeling the probability of success

The flight of the space shuttle Challenger

Statistical analysis in discrimination lawsuits

Predicting bankruptcy in the telecommunications industry

News flash! Smoking makes you live longer!

The sinking of the *Titanic*

What makes a top university top?

MYTHS ABOUT DATA ANALYSIS

1. *The results of a data analysis hinge on the statistical significance of hypothesis tests.*

Hypothesis tests are a useful tool to help determine what is going on in a data set, but they have no inherent superiority over other tools, such as graphical methods. Hypothesis tests can give misleading results when samples are small, when samples are very large, and when assumptions being made do not hold. **Don't fall in love with the number .05 — it is not a magic number!**

2. *There is a single correct way to analyze a given data set.*

There are many different ways to analyze a typical data set, each with their own strengths and weaknesses. Usually any reasonable analyses will end up with similar results and implications. **There is more than one path to the summit!**

3. *When you come to a point in your analysis where you have to make a decision, you only can choose one possibility and follow it until you're done.*

Good data analysis is a process of following up leads that often reach dead ends. If you're not sure what path to take at a given point, try both paths and see what happens — the only thing you lose is a little time. The answer to the question “I'm not sure if this will help; what should I do?” is always “Try it and see.” **Any choices you make that you can justify are okay, as long as you tell people what you are doing.**

4. *The goal of an analysis is to ultimately come up with a model that has the strongest measures of fit possible.*

There is only one goal in any data analysis — to uncover what is actually going on in the data. All data analytic decisions should be driven by that concern, **not** by whether they make the R^2 (or F , or t) larger. Don't succumb to “ R^2 envy” (“Ha ha! Mine is bigger than yours!”). **Good data analysis is very much like good detective work — its goal is not to verify our own beliefs, but rather to search for the truth.**