Department of Epidemiology and Biostatistics, McGill University

Information concerning course 513-678 "Analysis of Multivariable Data"

Monday, June 2 to Friday June 27, 2003

James Hanley, Instructor

This material revised 2 June, 2003

Timetable/Location/Facilities:

This is a 36 hour, 3 credit course.

With 3 exceptions*, the class will be held 1.30-4.45 in Room 409 of the McIntyre Medical sciences Building

 *June 2nd: Class on   begins at 2:00 p.m.   *June 13 and 23 :Class will be in Purvis Hall, Room 49*

Room 409 is equipped with Windows-based PC's for students and an instructor's PC with screen projection.

Prerequisites:

I will assume that everyone has had a first course in statistics, dealing with descriptive statistics; expectation and variance of random variables; the concepts of statistical tests/significance levels/confidence intervals; the application of these to one and two sample procedures for means (t and z distribution) and proportions (z and/or chi square distributions); and the beginning of simple linear regression.

These are the topics covered in the department's first course (607 - Inferential Statistics given in the Fall, or the two 2-credit sequence in May and June) and in almost every other first course in statistics given anywhere. If it has been a while since you have taken a course like this, I urge you to brush up on these topics. The text used in 607, Moore and McCabe's excellent "Introduction to the Practice of Statistics", covers all of them. It has a particularly good introductory chapter (2: Looking at Data: Relationships) which *everyone* would do well to read and fully digest before coming to this multivariate course. This chapter and the one entitled Inference for Regression (chapter 9 in edition 2, chapter 10 in edition 3) will be the point of departure for us. Indeed Chapter 10 in the 3rd edition is a very good introduction to multiple regression.

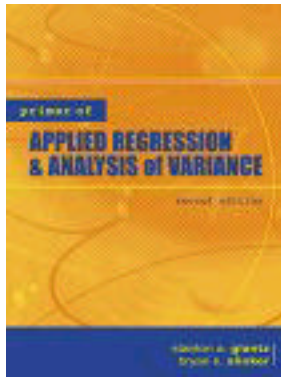Before the sessions on logistic regression we will review the log and exponential functions.

E-mail me if you are unsure whether you meet the pre-requisites.

22262

Content:

As you will see from the tentative agenda below, the first 2/3rds of the course will deal with "regular" multiple linear regression, and the last 1/3 or so with logistic regression. The 1st year [1996], we offered one session on survival analysis but we found only a few students wanted something this specialized, and in any case there wasn't enough time to really do it justice. We will determine by a survey of registered students how many would be interested in one session this year.

Text for the course:

For the first three weeks I plan to follow closely the (new) second edition of the text "Primer of Applied Regression Analysis & Analysis of Variance" by GLANTZ (Stanton A Glantz) and SLINKER (Bryan K Slinker)

Picture and info from amazon.com ..

> Hardcover - 949 pages 2nd edition (October 25, 2000)
> McGraw Hill Text; ISBN: 0071360867 ; Dimensions (in inches): 1.58 x 9.57 x 7.56
> amazon.com Price: $79.00 ($US)

This text will be available through McGill's bookstore (librarie) on McTavish Street (tel.398-7444) at a cost of approximately $125 $CDN before taxes.

This is the fourth time that we will use the Glantz and Slinker text (I have used Glantz introductory text a Primer of Biostatistics back before the Moore and McCabe days).

Four years ago there were such loud complaints from the 678 class about the 3rd Edition of the text "Applied Regression Analysis & Other Multivariable Methods" by Kleinbaum, Kupper, Muller and Nizam, Duxbury Press, Belmont Ca, 1998) that I do not plan to use it again. The KKMN text that has been bought by our regular students for the last 16 or so years in the department's full length

regression course 621 "Data Analysis I" but on careful questioning, most admit to me that they stay pretty close to Michal Abrahamowicz's course notes.

I have also asked the bookstore to have for sale Logistic Regression -- by David G. Kleinbaum, Mitchel Klein; Hosmer and Lemeshow's text Applied Logistic Regression; and Modelling Binary Data, Second Edition by David Collett (paperback) for those of you who will want to use logistic regression extensively in your future work. (You can check these out on amazon.com).

Many of the <u>principles </u>of multiple regression (coding of categorical variables, confounding, collinearity, stepwise regression, tests of fit, over-optimism and cross-validation) are the same whether we are dealing with regression methods for measured, counted or "survival- type" outcomes, so I would suggest you wait and see before buying a text specifically for the logistic regression material (Glantz and Slinker have a chapter on this same topic).

<u>Computing:</u>

The course is short; the material is new, calculations need to be done on a machine bigger than a hand calculator, and you need to gain practical experience in the analysis of real datasets. Thus, we need to "hit the ground running" and move forward fast, without getting hung up on mechanical problems of hardware/software. We will use the SAS program for all of our illustrations and you will be able to do the same in the teaching room during part of the class. If you need to, we can set up an account for you on our departmental server in Purvis Hall, so that you can log into it and run SAS, the Web browser, e-mail etc.. from one of the six computers in room 15 of Purvis Hall.

SAS is the package that the department supports and that our regular students are familiar with. Unless you already are using a comparable package (SPSS, Stata), we recommend that you too use SAS for your work in the course. It will help considerably if you can have some practice with SAS before the course begins. IF YOU WILL WANT TO USE IT IN YOUR LAB OR HOME COMPUTER, YOU NEED TO CONTACT US <u>SOON </u>ABOUT DOING SO --DON'T LEAVE THIS PART TO THE END!! We can, if students would like, arrange training sessions in our computer lab the week or two before the course and we will of course use some of the class time to do computer work. The exercises given out the very first day will involve use of the computer.

By the way, SAS now has one feature (Interactive Data Analysis, located under the "Globals/Analyze" menu in version 6 and "Tools/Analysis" in version 8) that makes it a lot less forbidding than previously. This feature allows the user to enter and view data in a spreadsheet format, and to carry out 'point and click analyses' rather than by typing commands into a program. Make sure if you arrange to put SAS on your computer that you install the INSIGHT module that

allows the graphical interface. This interface is available for a limited number of analyses, but they are exactly the ones we will be doing!

Format:

Each session of the course will devote some time to theory and some to application. Given that the text is quite well-organized and readable, we will tend to use transparencies and annotation tied to the text. Additional notes, diagrams datasets, etc, will be available on the class website.

Because it is counterproductive to work alone when learning to use a computer or computer package, feel free to form groups of 2 or 3 to work on the computing for the assignments and project. But you must hand in separate assignments, and your own annotation of any printouts.

We will have plenty of datasets to give out as assignments. Several of them, with the description, and the basic SAS program to read them, are already on the course web page

    http://www.epi.mcgill.ca//hanley/c678/

If you are registered for this course, e-mail me and I will give you the password to get into this web page. My email is James.Hanley@McGill.CA ;my telephone number 398 6270 and room no. 44 in Purvis Hall (top floor).

I cannot promise to have time to go over all of the intricacies of all the datasets students might bring to us during the course. However, I will try my best, and so I encourage you – if you have one --to bring a dataset of your own for analysis during the course (remember though that "data-cleaning" often takes a majority of the time spent on analysis, and so I urge you to come with the data already cleaned up).

Assessment:

The assessment will be based on homework assignments (60%) (given at each session) and a final project based on a dataset (40%).

Lastly, to get the most out of this course, it will take some serious time commitment on your part. Those of you who know me from previous years in 607 know that I work you very hard.

Course 607 covers a lot of statistical concepts such as variances, standard errors, p-values, etc., that are difficult for nonstatisticians to have a good intuition for or explain accurately in technical or nontechnical language. Fortunately, in a way, all of that is behind you at this stage, and the

"products" you will be producing during the course are much closer to what you need in your work and read in your literature. Moreover, you will get the computer to do the calculations for you. However, as with any of this material, it is only by hands on experience with real applications that you will master the concepts and principles involved.

I am very much looking forward to a great course with all of you and so urge you to get to it right now by:

So, to get ready...

- "brush up" on the prerequisites;

- scan the literature in your area to see how and why multivariate analysis is used;

- start to read the Glantz text (or KKMN if you think the comments 4 years ago were outliers) ;

- visit the website!

- IF YOU WHO HAVE NOT USED A SPECIALIZED COMPUTER STATISTICAL PACKAGE, START NOW TO LEARN! (the course is about data analysis, not *computing per se*)

    Every one step that you can practice now means more time to be able to spend learning multivariable analysis rather than computing. Don't be shy to contact me ahead of time about installing SAS on your computer (or setting up an account with us so you can use SAS in our Purvis Hall Lab). There is a SAS- INSIGHT Primer (27 pages) on the course web page..

    Fortunately, with each passing year, the amount and quality of the online help, and tutorials, for SAS and other computing packages keeps getting better. I put links to these at the top of the course 678 course page.