HRM723: Regression Analysis
January – April 2023

Coordinator: Dr. Lauren Griffith
MIP Suite 309a
☎: 905-525-9140, ext. 21416
E-mail: griffith@mcmaster.ca

Note: preferred method of contact is by email.

Objectives:

This is a second level course in statistics concentrating on several modeling techniques. There will be four main topic areas: 1) simple and multiple linear regression; 2) logistic regression, 3) survival data analysis, and 4) multi-level analysis. All of these use various forms of regression models.

Students will be required to carry out an analysis on a data set of their own choosing, using one or more of these techniques; this work will form a major basis for course evaluation. There are weekly assignments based on other data sets. Examples of uses of regression models in the literature will also be discussed in class.

Prerequisites: HRM 702, or equivalent.

Time and Place:

The standard time for the course is on Tuesdays from 1:00pm to 4:00pm. Tutorial sessions will generally be held from 1:00-2:15 and lectures from 2:30-4:00. Problem sets for tutorial session will be posted prior to each class. The lectures will be held MDCL 3023. Break-out sessions will take place in MDCL 3502, 3503, and 3505 (3 groups). Quizzes will be administered in MDCL 3023.

Format:

Between classes, you will work on one of several problems, which together cover all the concepts in the course. To get the most out of these assignments you should work on them at home and bring your completed work to class. When approaching each assignment, I encourage you to consider the question that need to be answered, the type of data you are dealing with, what comparison(s) you want to make, etc. before you select a suitable regression model to approach each problem.

Small **student-led** tutorial groups will meet from 1:00-2:15 in breakout rooms MDCL 3502, 3503, and 3505 to discuss how each person approached the assigned problem set, both conceptually and with the computer software, and the interpretation of the results. This will require that at least one person in each small group brings a laptop. Each week one or two students will be responsible for leading the tutorial session. There will be a faculty tutor assigned to each group who can help to clarify concepts. There will also be time during the tutorial sessions to discuss issues around your final projects with the tutor. After a short break, the full class will meet at 2:30 in the main lecture room to consolidate the findings from the tutorial sessions and discuss any additional issues. This will be followed by a lecture which will provide the background to undertake the next week's problem set.

This is a hybrid of the problem-based learning (PBL) approach in which solving problems in a small group setting is used as a vehicle for learning. The lectures will provide the tools necessary to solve the problems and will be structured to allow time to address "matters arising" from previous weeks.

On **Feb 7**, **Mar 7**, and **Mar 28**, there will be a 75 minute quiz instead of the small-group sessions. The quizzes will be administered in MDCL 3023.

Finally, at the end of the course you will be required to present your own analysis of a dataset of your own choosing, preferably arising from your own research or that of your supervisor.

Resources:

I hesitate to make too many recommendations, as I have found that statistics textbook preferences are very personal. Still, here are some ideas about resources that you could consult. There are a number of resources available for you.

1) The lectures, data sets, exercises and literature examples. These are not intended as a blueprint to indicate exactly what is and is not included in the course; instead, they are one of several resources you can use to try to understand the concepts of the course. I approximately follow the topics as given in the lectures, but there will be flexibility to respond to the needs and interest of the group.

2) Textbooks.

The main course textbook dealing with **linear and logistic regression and ANOVA** topics is

"*Applied Regression Analysis and other Multivariable Methods*" 5th Edition, by Kleinbaum, Kupper, Nizam and Rosman, 2014 (KKNR).

Purchase of this book is optional but highly recommended. Many textbooks cover similar ground. If you already own an intermediate level book, you need not necessarily rush out to buy another. Some of the exercises are taken from KKNR.

For **logistic regression**, I recommend:

"*Applied Logistic Regression*": DW Hosmer and S Lemeshow, 3rd Edition, (Wiley 2013). This is the current "gold standard" text on this topic.

For **survival analysis**, I recommend:

"*Survival Analysis: A Self-Learning Text*". David G. Kleinbaum, 3rd Edition (Springer, 2012) or "*Applied Survival Analysis: Regression Modeling of Time-to-Event Data*" DW Hosmer, S Lemeshow, and S May, (Wiley, 2008).

For **multilevel analysis**, I recommend:

"*Multilevel Statistical Models*". H Goldstein, 4th Edition (Wiley, 2011) or
"*Applied Mixed Model Analysis: A Practical Guide*", JWR Twisk, 2nd Edition (Cambridge University Press 2019).

3) Software manuals and websites for programs such as SPSS, Stata, or SAS also may have some background on many of the topics covered in this course.

   Great site from University of California: http://www.ats.ucla.edu/stat/

4) HRM Stats Resources under "Continuing" Section on Avenue to Learn provides background and examples for several types of analyses completed in SPSS, SAS, and R

None of these resources is sufficient by itself. Like the real world, you are expected to approach problems using multiple methods.

Computing:

Because of the diversity of computing environments available to students, and their previous experience with various software programs, the current position of the HRM program is that we do not mandate the use of any specific package for this course. The exercises will be generally illustrated using SPSS and/or SAS, and the Kleinbaum text tends to use SAS, but suitable modification will usually allow one of the other major packages to do the same things.

SAS is big, powerful, and usually expensive! SPSS is cheaper and more popular, but the student version does have some analytic limitations, so sometimes people get frustrated when they find out it won't do exactly what they want, or if they can't understand the output. SAS and SPSS are available in the HRM student room. HRM students also have access to a free SAS download that they contact the program for (there is an agreement that they must sign and a download). Students can purchase the license for SPSS through titles bookstore.

In considering a package for this course, make sure you will be able to access logistic regression and survival analysis methods (preferably including life tables and the Cox proportional hazards model). Versions 10 and later of SPSS do logistic (and polytomous) regression. SAS also has a good logistic module. SAS and SPSS all have good capabilities for survival data analysis. If you anticipate having unusual analytic requirements that are not handled by routine packages, you may want to consider using a program such as Stata, S-Plus, or R; these require more programming skills, but do provide greater customisation and flexibility in the form of output you can generate.

If you are already expert with SPSS, SAS or other suitable program, don't switch. Most tutors have some general familiarity with each of SAS and SPSS, so they will be able to help you if you choose them. I can also provide some general support for other programs. I can't promise to solve all your problems, however, especially difficulties concerning installation and running of software on your system in particular. It is **your** responsibility to make sure you have access to adequate statistical software as soon as possible after the start of the course.

It is desirable to have a statistical package installed on a laptop that you can bring to tutorials.

Student Evaluation:

As previously mentioned, there are weekly assignments based on the data-based exercises. These assignments are there for your learning and generally will not be handed in. The exception is the logistic regression assignment which is due on **Feb 28** and will be marked. The learning objectives of each assignment will be discussed in class. Participation in the small groups and the in-class discussion will also count toward your final participation mark.

The final grade will be based on:

1. Student's participation in small groups and class and punctuality/attendance (5%).

2. Three multiple choice and short answer quizzes on **Feb 7**, **Mar 7** and **Mar 28**. Each quiz is worth 10% for a total of 30%. These will test your understanding of basic concepts and techniques.

3. One tutorial assignment will be handed in prior to the tutorial on **Feb 28** and marked. This assignment is worth 15%.

4. A final assignment at the end of term (45%). This mark has three components: A 1-page description of your dataset, research question and proposed analysis that will be due **Feb 14** (5%), a hand-in report (25%), an oral presentation (15%) will be due in the last three weeks of the course [details below]. I encourage you to analyze your own data set if possible.

5. Commenting on one student's final presentation and project and leading their discussion (5%)

The instructor reserves the right to modify elements of the course and will notify students accordingly either in class or on Avenue to Learn.

Final assignment

The final assignment on your own project consists of a hand-in report, of at most 6-8 double-spaced pages using 12 point font, and a class presentation of 15 minutes. The report can be accompanied by a maximum of 6 tables and/or figures in addition to the written text. Excess pages in the report will not be read. The presentation should be about 10 minutes, leaving at least 5 minutes for discussion. I would like you to apply one or more of the methods discussed in class to a problem of your own. The discussion will be led by one of the other students who has been randomly assigned to provide you feedback.

I encourage you to consider the following in choosing your data:

- It should be non-trivial in terms of the number of variables available. Anything that only requires a t-test or a 2x2 table would not be sufficient.

- Conversely, it need not be big. You don't necessarily need a thousand cases to make the design and analysis complex and interesting. Many interesting analysis problems come from small studies with several experimental factors.

- Larger databases may contain lots of missing data, which can itself be an issue for the analysis. Missingness does not rule out use of a big data set, but the reasons for missing data and potential biases would need to be explored.

- Please use data where the dependent variable is interval or ratio for linear regression, binary if you are using logistic regression, or time-to-event for survival analysis. Multilevel analysis is also allowed.

Please start **NOW** to look for a suitable problem and database, and don't leave it till the last few weeks. Obviously I don't necessarily expect you to be able to analyse the data early in the course. But some time spent now to find the right kind of data will save more time and anguish later.

Some students elect to use statistical techniques that have *not* been directly covered in the course (examples include discriminant function analysis, questionnaire scaling, etc.). This is OK, because many of these methods are *based* on models that are included in the scope of the course. If you are inclined in this direction, we can provide some individualised support, but remember that you will be ultimately responsible for the appropriateness and validity of your analysis.

Projects will be presented to the class in the last three weeks of the course, and then the final written report handed a week after your oral presentation. I would encourage you to describe your alternative approaches to the analysis, and any problems you encountered in your presentation, since I have always found that everyone learns a lot from the experiences of others in the group.

The ability to write about statistical methods and results at a professional level is an important aspect of evaluation. You can NOT include computer output in your written report. All relevant aspects of the output should be summarized in tables and/or figures. I should be able to understand your approach by reading the statistical methods section of your paper. I may, however, ask for output if required to mark your paper. In general, for both the verbal and written versions of the projects, graphics are often a useful way to express results.

You may find filling in the following table helpful when thinking about your proposed analysis but it should not be handed in as your 1-page proposal due on Feb 14:

Objective	**Outcome**	**Predictors**	**Hypotheses**	**Sample Size**	**Method of Analysis**
Primary					
Secondary					

Your presentation will likely include the following:

a) Introduction - background to the health topic being analysed
b) The questions to be answered, hypotheses to be answered.
c) The data available.
[approximately 1 page]

d) The statistical methods
[approximately 1 page]

e) Descriptive data: e.g. distributions, plots.
f) The main analyses answering your questions.
[approximately 2 pages]

g) Interpretation of the results.
h) Conclusions/Discussion
[approximately 2 pages]

I will try to provide some opportunities to discuss your projects during the term. I have found it most profitable to discuss these issues with the whole class group. (Often several people may have encountered the same problems). You should have time to discuss your projects during the tutorial sessions as well.

During the last three weeks you will present your final project on one day. You will also be asked to lead the feedback discussion for another student during a separate week. These dates will be selected in advance and posted on Avenue.

<u>Schedule:</u>

Here is the approximate schedule of topics:

SESSION	**TOPIC**
1 – 4	Course Introduction; linear regression (Slides and KKNR Chapters 5-14, 16)
5 –7	Logistic regression (Slides and KKNR Chapters 22-23 or Hosmer and Lemeshow Chapters 1-8)
8 – 10	Survival Analysis (Slides and Kleinbaum Chapters 1-6)
11	Mutli-level Analysis (Slides)
12-14	Final Presentations

VERY IMPORTANT: This course is usually fully subscribed and there may be a waiting list. We try to accommodate all HRM students. If you have changed your mind about taking the course, please let the Graduate School or Kristina Vukelic know right away, so we can replace you with another student.

HRM policy on absenteeism

1. Any absence must be due to a reasonable excuse that is exceptional and out of the control to some extent of the student (illness, death in family, wedding, special exams etc).

2. One absence with an reasonable excuse is acceptable, two may be ok, but three would almost never be acceptable.

3. If you are absent you get 0 for participation on that day.

4. Greater than 2 absences must be approved by the HRM program coordinator. This includes an absence on the final presentation day on which you are not presenting or leading the discussion.

Notes: For information about academic dishonesty, ethical issues and students with disabilities, please see the School of Graduate Studies calendar.