MATH 60604A
Statistical Modelling

Juliana Schulz

HEC Montréal
Department of Decision Sciences

Fall 2025

About me

- Juliana Schulz
- Background in actuarial mathematics
  BSc and MSc at Concordia
  Experience primarily in non-life insurance
- 2018: PhD in statistics, McGill University
  Multivariate Poisson models based on comonotonic and counter-monotonic shocks
- 2018–2019: Postdoc in biostatistics, McGill University
  Doubly Robust Estimation of Optimal Dosing Strategies
- Assistant (2019–2024) / Associate (2024–) Professor, HEC Montréal, Department of Decision Sciences
- My research interests:
  multivariate statistics, dependence modelling, biostatistics (causal inference, optimal dynamic treatment regimes), actuarial mathematics (non-life insurance)

You?

- Name? Program?
- Why are you in data science / why are you taking this couse?

Course Objectives

- learn the statistical tools for data modelling in a regression framework
  - basic principles in inference and statistical modelling
  - linear models
  - generalized linear models
  - linear mixed models (for longitudinal/correlated data)
  - introduction to survival analysis
- learn the theory behind the statistical models
- concrete examples with applications; data analyses

Course objectives

- Statistical modelling and inference in the context of regression models
- Regression model:
  - Goal: **relate** a response variable of interest Y with one or more explanatory variables $X_1, X_2, \ldots, X_p$
  - Examples:
    - Y: salary, X_1: age, X_2: experience
    - Y: amount spent on purchase, X_1: payment method (cash, credit, debit)
    - Y: monthly energy consumption of a household, X_1: number of people living in the home, X_2: size of the home, X_3: average temperature during the month

Course objectives

- To establish a link between Y and $X_1, X_2, \ldots, X_p$, we will use basic regression modelling tools:
  - Linear regression models
    - simple linear regression
    - multiple linear regression
    - (t-test and ANOVA in regression framework)
  - Generalized linear models
    - logistic regression
    - Poisson regression
  - Linear regression models for correlated and longitudinal data
    - linear regression with random effects
  - Survival analysis
    - cox proportional hazards model
- Data applications using R

Explanatory versus predictive models

- In general, there are two main reasons we would be interested in modelling the relationship between a dependent variable Y and a group of explanatory variables $X_1, X_2, \ldots, X_p$:
  - Predictive model
    - Goal: make predictions
    - Allows us to make predictions on the value of Y for future observations of $X_1, \ldots, X_p$
    - Ex: predict the energy consumption for a home (Y) as a function of weather (X_1), number of inhabitants (X_2), house size (X_3)
  - Explanatory model (inference)
    - Goal: see how the the explanatory variables affect Y
    - Allows us to test research hypotheses concerning the variable effects
    - Ex: What is the effect of the payment method (X_1) on the total amount spent on purchases (Y)?

Explanatory versus predictive models

- In this course, the focus is on using regression models as explanatory models: this is statistical inference
  - Note that the models that we will see can also be used for predictions, although this is not the emphasis here.
  - Predictive models are the focus of MATH 60603(A) Statistical Learning.
- We can use the same tool for many purposes, but *how* we use it depends on our goal.
  - While a regression model can be used to answer both "predictive" and "inferential" type questions, the underlying objective (prediction vs. inference) will change our approach for analyzing data (how to define the model, how to validate / evaluate the model, etc.)
- Predictive models:
  - Machine learning models generally focus on prediction
  - "Black box" models: interpretation is often difficult, which can be problematic in many practical situations

Explanatory versus predictive models

- Explanatory models:
  - We're interested in explaining the relations between variables: quantifying the effect of a variable on another, evaluating whether this relation is "significant", etc.
  - Interpretability is very important.
- Regression models are frequently used for explanatory purposes
  - allow for interpretation
  - allow to measure and test the variable effects
- Statistical inference is a very important part of data analysis
  - it allows us to infer; to generalize the results based on an analysis of sample data to a bigger population

Overview of course material

- Throughout the course, the focus is on the case where there is a response variable of interest Y and one or more explanatory variables $X_1, X_2, \ldots, X_p$.
- The overall objective of this course is to understand how to evaluate the effects of the variables $X_1, \ldots, X_p$ on the variable Y - this is essentially what regression models do
- All of the chapters in this course will revolve around this general problem, in different contexts and for different types of variables $X_1, \ldots, X_p$ and Y
- All of the chapters in the course can be expressed in the form of a regression model (relating the explanatory variables $X_1, \ldots, X_p$ to the response variable Y)

Overview of course material

Type of variable Y	Independent Observations	Method
Continuous	Yes	**Simple linear regression (chap 2 part 1)** **Multiple linear regression (chap 2 part 2)** **Special cases: t-test and ANOVA (chap 2 part 3)** **Models for survival data (chap 6)**
Continuous	No (ex : longitudinal study)	**Regression with random effects (chap 5)**
Binary	Yes	**Logistic Regression (chap 4)**
Count	Yes	**Poisson Regression (chap 4)**

Overview of course material

Chapter 1 - Introduction / Review

- Quick review:
  - Probability
  - Random Variables
    - some basic distributions (discrete, continuous)
    - moments
  - Inference
    - basic notions
    - sampling, estimation
    - hypothesis testing
    - confidence interval

Overview of course material

Chapter 2 - Linear regression

- Part 1: Simple Linear Regression
  - How to measure the (linear) effect of a single explanatory variable X on a (continuous) response variable Y
    - Ex: Modelling salary (Y) as a function of years of experience (X_1).
- Part 2: Multiple Linear Regression
  - How to measure the (linear) effects of several explanatory variables $X_1, \ldots, X_p$ on a (continuous) response variable Y
    - Ex: Model salary (Y) as a function of years of experience (X_1), schooling (X_2), and sex (X_3).

Overview of course material

Chapter 2 - Linear regression

- Part 3: Special cases
  - t-test: comparing 2 population means
    - equivalent to fitting a linear regression model with a single binary explanatory variable
    - ex: comparing the average salary (Y) in Quebec City vs. Montreal (X) $\rightarrow$ testing whether there is a relation between X and Y is equivalent to testing whether the mean salary in the 2 groups (Quebec vs. Montreal) differs
  - ANOVA: comparing population means in $k > 2$ groups
    - equivalent to fitting a linear regression model with a categorical explanatory variable (with k levels)
    - ex: comparing the average salary (Y) in 4 cities: Quebec City, Montreal, Ottawa and Toronto (X) $\rightarrow$ testing whether there is a relation between X and Y is equivalent to testing whether the mean salary in the 4 cities differ.

Overview of course material

Chapter 3 - Maximum likelihood estimation

- Concept of likelihood function
- Maximum likelihood estimation
- Likelihood-based tools

Overview of course material

Chapter 4 - Generalized linear models

- Intro to Generalized Linear Models
  - Introduce the class of generalized linear models which can accommodate different forms of response variables Y (includes linear regression as specific case)
- Logistic regression
  - Case where Y is binary (i.e. $Y \in \{0, 1\}$)
  - Ex: logistic regression: modelling the presence (or absence) of exmployees as a function of the characteristics of the day
- Poisson regression
  - Case where Y is a count variable (i.e. $Y \in \{0, 1, 2, \ldots\}$)
  - Ex. Poisson regression: modelling the number of accidents at intersections as a function of road conditions.

Overview of course material

Chapter 5 - Linear mixed models

- Correlated and longitudinal data:
  - data with dependent observations Y
    - ex: longitudinal study - measurements taken over time on same individual
    - ex: correlated data - measurements taken from subjects which are not independent ("grouped" data)
  - notion of correlation structure
- Linear mixed models:
  - extension of linear regression (still in case where Y is continuous) that accounts for dependence
  - ex: modelling the *progression* of salaries (Y) as a function of age (longitudinal study)

Overview of course material

Chapter 6 - Introduction to survival analysis

- Survival data:
  - data are the time until the occurrence of an event of interest
    - ex: time until a customer cancels their gym membership
    - ex: time until a light bulb burns out
  - notion of censored data
- Survival function:
  - Kaplan-Meier Estimator (non-parametric)
  - Comparing survival curves (log-rank test)
- Cox Proportional Hazards Model:
  - particular regression model (semi-parametric) for survival data

Prerequisites

A first course in probability/statistics covering the following notions:

- basic probability
- random variables:
  - discrete distributions (Bernoulli, binomial, Poisson)
  - continuous distributions (uniform, exponential, Normal)
  - moments (expectation, variance)
- statistical inference:
  - sampling and estimation
  - hypothesis tests
  - confidence intervals
- Don't panic - review material will be provided, and we will go over these concepts again in the context of regression models.
- For more help: the Mathematics and Statistics Help Centre
  https://www.hec.ca/en/mshc/index.html

Working in R

- We'll be working in R throughout the semester to carry out various statistical analyses.
- R:
  - R is a free software environment for statistical computing and graphics

    https://www.r-project.org/
  - examples in R (sample code will be provided throughout the semester)
- R studio:
  - RStudio is a (free) integrated development environment (IDE) for R

    https://www.rstudio.com/products/rstudio/
- Download:
  1) Download R https://cran.rstudio.com/
  2) Download RStudio (RStudio Desktop: free, open source edition) https://www.rstudio.com/products/rstudio/download/

Working in R

- Within RStudio:
  - R code (R file): programming language used to carry out statistical computations, create graphics, etc.
  - RMarkdown (RMD file): tool available in RStudio, convenient for writing reports while incorporating R code and output
  - You will be required to use R (and RMarkdown) throughout the semester!

Working in R: resources

- Tutorial available on ZoneCours
  - Introduction to R parts 1, 2, 3, 4, exploration
    - this tutorial consists of a series of documents providing a brief introduction to R, RStudio and RMarkdown (along with corresponding RMarkdown files, R code, and datasets to guide you through)
  - These documents should provide a sufficient introduction so that you feel comfortable working with R / RMarkdown.
- Additional resources:
  - Centre d'aide en mathématiques et statistique (CAMS)
    - Various (online) workshops available:
      `https://hec-ca.libcal.com/calendar/cams`
    - An introduction to R: (September 9, 15:30–18:30)
    - online, must register, limited places!
- NOTE: the objective of this course is not to become expert R coders, rather we will be using R to carry out data analyses

Schedule

- 2 sections:
  - Wednesdays 3:30PM – 6:30PM (English)
  - Thursdays 12:00PM – 3:00PM (French)
- 12 weeks of class + midterm + final exam
- 6 chapters:
  1) introduction/review
  2) linear regression: simple, multiple
  3) maximum likelihood estimation
  4) generalized linear models: logistic regression, Poisson regression
  5) linear mixed models
  6) survival analysis
  7) (review)

Schedule

Tentative schedule:

Lecture	Date	Material	Exam / Project
1	27-Aug	Ch 1 introduction	
2	3-Sep	Ch 2 linear regression	
3	10-Sep	Ch 2 linear regression	
4	17-Sep	Ch 2 linear regression	
5	24-Sep	Ch 3 likelihood based inference	
6	1-Oct	Ch 4 generalized linear models	Project: part 1
7	8-Oct	Ch 4 generalized linear models	
	15-Oct	NO CLASS	
	23-Oct		**Midterm exam**
	29-Oct	NO CLASS	Project: part 2
8	5-Nov	Ch 5 linear mixed models	
9	12-Nov	Ch 5 linear mixed models	
10	19-Nov	Ch 5 linear mixed models	
11	26-Nov	Ch 5 linear mixed models	Project: part 3
12	3-Dec	Ch 6 survival analysis / review	
	4-Dec		**Final exam**

** Please note that this is an approximate timeline and may change*

Course Format

- Classes will be given
  - **in person**
  - (no Teams/Zoom access, no recordings)
- Some class rules:
  - I expect you to participate
  - don't hesitate to ask questions (raise your hand or simply speak up)
  - please stay in your class section (i.e. don't attend the other section lectures)

ZoneCours & Teams

- ZoneCours: for all important information
  - detailed course outline
  - class material (class notes, exercises, instructions, etc.)
- Teams: MATH 60604A - A25
  - join with code `fkte3ia`
  - can chat amongst peers here, ask questions, etc.

Evaluations

- Project (team work): 30%
  - consists of 3 parts, due dates spread out throughout the semester
- Midterm exam (individual work): 30%
  - in person, written exam
- Final exam (individual work): 40%
  - in person, written exam
- Note that plagiarism, in any form, will not be tolerated (grade of 0)!
  - make sure you know what plagiarism is, particularly with regards to AI

Use of Generative Artificial Intelligence (GAI)

Taken from Zone Cours:

Generative Artificial Intelligence (GAI)

Students must familiarize themselves with the specific instructions for each course evaluation in order to validate whether or not they are permitted to use GAI.

When its use is permitted, students are responsible for ensuring that they comply with the applicable rules, particularly with regard to citation and mention of sources. Any use of GAI that does not comply with the applicable rules constitutes an academic offence.

Where its use is prohibited, it is also an academic offence to use GAI.

Any academic offence is likely to give rise to procedures and sanctions provided for in regulations on intellectual integrity.

Each student is responsible for informing himself or herself and adopting good practices related to intellectual integrity. To find out more, consult the following resources:

- Intellectual integrity at HEC;
- Ways to avoid plagiarism;
- How to cite your sources.

Use of Generative Artificial Intelligence (GAI)

- Exams (midterm + final): prohibited
- Project: prohibited, except for specific uses
  - students must create the original draft of all components of the project (including all statistical analyses, code, and accompanying text)
  - students must keep a copy of these original drafts, which can be provided upon request
  - the use of GAI tools is permitted to enhance / edit the work (including the text and code) from a review perspective:
    - GAI may be used to assist in debugging and/or cleaning code, but may not be used to create the code from scratch
    - GAI may be used to correct and/or improve the syntax of the text, but may not be used to create the text from scratch
  - any use of GAI must be appropriately cited, along with an accompanying appendix which summarizes the prompts / querries used with GAI tools, as well as an explanation of how the content generated by GAI tools was used / adapted in the submitted work
  - note that students are fully responsible for the content of the work submitted for evaluation, and as such, an error produced by a GAI tool will be treated as such in the correction - the fact of having obtained information from a properly cited source does not excuse any errors produced
  - failure to follow these rules will be considered an academic offense

Use of Generative Artificial Intelligence (GAI)

- Caution: an analogy of using GAI to code in R (`https://fediscience.org/@andrew/112599267984824456`)

I am **not** anti-ChatGPT. I use LLMs like ChatGPT and GitHub Copilot all the time in my own work (GitHub Copilot in particular is *really really good* for code—it's better than ChatGPT, and free for students). These tools are phenomenal resources ***when you already know what you are doing***.

However, using ChatGPT and other LLMs when *learning* R is actually really detrimental to learning, especially if you just copy/paste directly from what it spits out. It will give you code that is wrong or that contains extra stuff you don't need, and if you don't know enough of the language you're working with, you won't understand why or what's going on.

Using ChatGPT with R requires a good baseline knowledge of R to actually be useful. A good analogy for this is with recipes. ChatGPT is really confident at spitting out plausible-looking recipes. A few months ago, for fun, I asked it to give me a cookie recipe. I got back something with flour, eggs, sugar, and all other standard-looking ingredients, but it also said to include 3/4 cup of baking powder. That's wild and obviously wrong, but I only knew that because I've made cookies before. I've seen other AI-generated recipes that call for a cup of horseradish in brownies or 21 pounds of cabbage for a pork dish. A few weeks ago Google's AI recommended adding glue to pizza sauce to stop the cheese from sliding off. Again, to people who have cooked before, these are all obviously wrong (and dangerous in the case of the glue!), but to a complete beginner, these look like plausible instructions.

Use of Generative Artificial Intelligence (GAI)

- Useful links:
  - https://www.hec.ca/en/generative-artificial-intelligence/index.html
  - https://www.hec.ca/en/students/support-resources/generative-artificial-intelligence/use-responsibly/index.html

Project

- You will be analysing BIXI data from the 2024 season.
- The project is divided into 3 parts (tentative deadlines):
  - part 1: linear regression models (due October 2, 11:55 PM)
  - part 2: generalized linear models (due October 29, 11:55 PM)
  - part 3: linear mixed models (due November 27, 11:55 PM)
- Details:
  - counts towards 30% of your final grade
  - team work (groups of 3* students) – notify me of your teams as soon as possible!
- More detailed instructions to come...

Exams

- Midterm exam (30%):
  - October 23 from 12:00 –15:00
  - in person, written exam
  - covers roughly chapters 2 – 4
  - a single-sided sheet of handwritten notes will be permitted
  - calculator (non-programmable HEC approved) permitted
- Final exam (40%):
  - December 4 from 18:30 – 21:30
  - in person, written exam
  - covers all course material
  - a double-sided sheet of handwritten notes will be permitted
  - calculator (non-programmable HEC approved) permitted

Class Format

- Course material:
  - class notes all you need (slides + examples in R)
  - references listed on the course outline are supplementary material and are NOT required
- Alternating between theory and applications:
  - theory behind models
  - data illustrations
  - examples in R
- Breaks

- Do not hesitate to ask questions during class!

Managing expectations: professor

- All class material will be posted on Zone cours

  (careful: sometimes there will be revisions to the class notes, always be sure to have the most updated version!)

- I will do my best to answer all your question
  - in class (raise your hand, or just call out)
  - by email (if the answer is short)
  - by appointment (it's often helpful to discuss problems in person, even if virtually)
  - remember... there are no stupid questions!
- I will do my *best* to correct course work within a timely manner (approx 2 – 3 weeks)
- Don't hesitate to give me feedback throughout the semester so I can try to adapt my approach to best suit the class' needs
  - but please remember, you are not alone in this class!

Managing expectations: students

- School's rules:

  According to the School's rules, students are expected to attend classes (or course activities). Instructors are not required to provide any additional help or adapt courses or evaluations due to a student's absence.

- Come to class prepared:
  - go over material from the previous class
  - bring your laptop
  - download material from ZoneCours (class notes, examples, R code, datasets, etc)
  - follow along in class
  - do the exercises
  - above all... PARTICIPATE in class and don't hesitate to ASK QUESTIONS

Support & Ressources

- HEC Montréal support and ressources for students:

  https://www.hec.ca/en/students/support-resources/index.html