Welcome to DSCI 562: Regression II

Contents

This course focuses on exploring regression techniques beyond Ordinary Least-squares (OLS). Some specific inquiries are the following:

- What if the response is still continuous but non-negative, or our observations are not independent anymore?
- Maybe the response is binary, a count, or categorical.
- Moreover, what if we have censored data?

- In many other cases, we might be interested in modelling a response different from a conditioned mean. Different Data Science-related projects will require making inference on different conditioned response quantiles.

We will cover useful extensions to classical linear regression: generalized linear models (GLMs), mixed-effects, local, survival, and quantile regression, and techniques for dealing with missing data.

High-Level Goals

By the end of the course, students are expected to:

- Describe the risk and value of making parametric assumptions in regression.
- Fit model functions that represent probabilistic quantities besides the mean.
- Identify situations where OLS regression is sub-optimal, and apply alternative regression methods that better address the situation.

Teaching Team

Position	Name	Slack Handle	GHE Handle	Section
Lecture/Lab Instructor	Payman Nickchi	`@Payman Nickchi`	`@pnickchi`	1
Lecture/Lab Instructor	Alexi Rodríguez-Arelis	`@Alexi`	`@alexrod6`	2
Teaching Assistant	Jared Connoy	`@jaredconnoy`	`@connoyja`	1
Teaching Assistant	Tony Fong	`@Tony Fong (TA)`	`@tf4ong`	1
Teaching Assistant	Meltem Omur	`@Meltem Omur (TA)`	`@Momur17`	1
Teaching Assistant	Atabak Eghbal	`@Atabak`	`@ateghbal`	2
Teaching Assistant	Samir Damji	`@Samir Damji`	`@samir10`	2
Teaching Assistant	Haley Oleynik	`@Haley Oleynik (TA)`	`@holeynik`	2

Lecture Schedule

This course occurs during **Block 4** in the 2024/25 school year. The course notes can be accessed **here**. Typically, you should review these notes before each lecture.

Lecture	Topic	Optional Readings
1	**Generalized Linear Models: Link Functions and Count Regression**	- Chapter 1 (*Review of Multiple Linear Regression*): BMLR (**optional, only as a review**) - Chapter 6 (*Logistic Regression*): BMLR (**optional, only as a review**) - Chapter 5 (*Generalized Linear Models: A Unifying Theory*): BMLR (**short chapter**) - Chapter 4 (*Poisson Regression*): BMLR
2	**Generalized Linear Models: Model Selection and Multinomial Logistic Regression**	- Chapter 2 (Beyond Least Squares: Using Likelihoods): BMLR - Faraway: 5.1 (Multinomial Logit Model)
3	**Generalized Linear Models: Ordinal Logistic Regression**	- Faraway: 5.3 (Ordinal Multinomial Responses)
4	**Linear Mixed-Effects Models**	- Initial motivation (ISL): 3.3.3 Potential Problems (*2. Correlation of Error Terms*) - Chapter 8 (*Introduction to Multilevel Models*): BMLR (from 8.1 to 8.8)
5	**Survival Analysis**	See Kleinbaum and Klein (2005) in Reference Material
6	**Local Regression**	- ISL 7.2 Step Functions - ISL 7.6 Local Regression

Lecture	Topic	Optional Readings
7	**Quantile Regression**	- Fahrmeir (2013) 10. Quantile Regression (except 10.2.2 Bayesian Quantile Regression)
8	**Missing Data**	See van Buuren, S. (2012) in Reference Material

See the lecture learning objectives for a detailed breakdown of lecture-by-lecture learning objectives.

Regression Mind Map

Here is a mind map we created to summarize all regression models to be covered in this course.

Deliverables

This is an **assignment-based course**. The following deliverables will determine your course grade:

Assessment	Weight
Lab Assignment 1	12%
Lab Assignment 2	12%
Lab Assignment 3	12%
Lab Assignment 4	12%
Quiz 1	25%
Quiz 2	25%
Lecture Attendance (iClicker)	2%

> **Note:** A +1% final bonus mark will be granted to everybody if the class reaches a 75% response rate (or above) in the final teaching evaluations.

Lectures

Schedule

Refer to the MDS calendar.

Labs

Schedule

Refer to the MDS calendar.

Lab Topics and Due Dates

	Lab Topic	Due Date
1	Introduction to Generalized Linear Models (Lectures 1 and 2)	Refer to the MDS calendar.
2	Ordinal and Mixed-Effects Regression Models (Lectures 3 and 4)	Refer to the MDS calendar.
3	Survival Analysis and Local Regression (Lectures 5 and 6)	Refer to the MDS calendar.
4	Quantile Regression and Missing Data Imputation (Lectures 7 and 8)	Refer to the MDS calendar.

Lab Grade Computation

Once lab grades are published on Gradescope, you will see your **raw lab mark** m. This **raw lab mark** m is the grand total of your granted marks throughout the whole lab assignment. Now, if we add up **all the marks (non-challenging and challenging)** in the handout corresponding to all `rubric={...}`, this sum is what we call the maximum raw lab mark m_{100} to get 100% as a percentage lab grade. On the other hand, if we add up **the non-challenging marks** in the handout found in `rubric={...}`, this sum is what we call the raw lab mark m_{95} to get a 95% as a percentage lab grade.

By the end of the block, **once all lab marking is finished on Gradescope**, your raw lab grades will be transferred to **Canvas**. Then, in your **Canvas gradebook**, you will see these raw lab grades (`raw lab1`, `raw lab2`, `raw lab3`, and `raw lab4`). Finally, for each of the four labs, you will also see your final lab grades (`lab1`, `lab1`, `lab3`, and `lab4`). Let g be the final lab grade of a specific lab **as a percentage**; it will be computed as follows:

- If $m > m_{95}$, then $g = 95 + \left(\frac{m - m_{95}}{m_{100} - m_{95}} \times 5 \right)$.
- If $m \leq m_{95}$, then $g = \left(\frac{m}{m_{95}} \right) \times 95.$

Quizzes

Refer to the MDS calendar.

Office Hours

Refer to the MDS calendar.

Communication

We will use **Slack** as the main communication channel.

If you have any questions regarding the course content, lectures, labs, autograders, or any other course-related matters, we kindly request that you avoid direct messaging (DM) the instructor or TAs. Instead, please post your question on the DSCI 562 channel. This approach not only enables our TAs to respond promptly but also benefits other students who might have similar questions.

> **Response time:** We will try our best to reply to your inquiries as soon as possible during the normal working hours (9AM-5PM Mon-Fri). If you send us a message outside of regular working hours, please expect a response on the next working day.

Use of LLMs

LLMs, such as ChatGPT, can be helpful tools if we use them responsibly. In this course, students are permitted to use these tools to gather more information, review concepts, or brainstorm, and students must cite these tools if they use them for assignment. Having said all this, it is **not** permitted to write any given assignment via copying and pasting AI-generated responses.

Reference Material

- Agresti, A (2013). *Categorical Data Analysis*, John Wiley & Sons, Incorporated. ProQuest Ebook Central.
    - The e-book is available through the UBC Library. You can obtain a PDF copy with your CWL account. This book is helpful for generalized linear models with discrete responses.
- Collett, D. (2003). *Modelling Binary Data (2nd ed.)*. Chapman and Hall/CRC. https://doi.org/10.1201/b16654
    - The e-book is available through the UBC Library.
- Fahrmeir, L. (2013). *Regression Models, Methods and Aplications*. Springer Berlin Heidelberg.
    - The e-book is available through the UBC Library. You can obtain a PDF copy with your CWL account.
- Faraway, Julian J. (2005). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, CRC Press LLC. ProQuest Ebook Central.
    - The e-book is available through the UBC Library. You can obtain a PDF copy with your CWL account. This book is great for learning how to work within the R environment with the models we will be working on. Its approach is essentially practical.
- Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press.
    - The physical book is available through the UBC Library. This book is pretty useful and practical as introductory material.
- Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Publising Company, Incorporated.
    - The e-book is available through the UBC Library.
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An Introduction to Statistical Learning: with Applications in R*. Springer Publishing Company, Incorporated.
- Kleinbaum, D. G. and Klein, M. (2005). *Survival analysis : A Self-Learning Text*. Springer.
    - The e-book is available through the UBC Library. This book is a good start for Survival Analysis:
        - **Chapter 1 (Introduction):** Introduction to Survival Analysis (I), Censored Data (II), Terminology and Notation (III).

        - **Chapter 2 (Kaplan-Meier Curves):** Review (I), Example of Kaplan-Meier Curves (II), General Features of Kaplan-Meier Curves (III), Confidence Intervals for Kaplan-Meier Curves (VII and VIII).
        - **Chapter 3 (Cox Proportional Hazards Model):** Example of Cox Proportional Hazards Model (I), Formula of Cox Proportional Hazards Model (II), Why the Cox Proportional Hazards Model is Popular (III), Estimation of the Cox Proportional Hazards Model (IV).
        - **Chapter 7 (Parametric Model):** Overview (I), Relationship Between the Probability Density Function with Hazard and Survival Functions (II), Weibull Example (IV).
- Rousseeuw, L. P. J. and Leroy A. M. (2003). *Robust Regression and Outlier Detection*. Hoboken, NJ : Wiley-Interscience.
    - The e-book is available through the UBC Library. You can obtain a PDF copy with your CWL account.
- Roback, P. and Legler, J. (2020). *Beyond Multiple Linear Regression*.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.
    - The e-book is available through the UBC Library. You can obtain a PDF copy with your CWL account.
- van Buuren, S. (2012). *Flexible Imputation of Missing Data*:
    - **1.1 The problem of missing data.**
    - **1.2 Concepts of MCAR, MAR and MNAR.**
    - **1.3 Ad-hoc solutions.**
    - **1.4 Multiple imputation in a nutshell.**

Recommended Course Reviews

This course is entirely taught in `R` (we will follow the `tidyverse` style guide) with a reasonable mathematical and statistical basis. We strongly recommend reviewing the following courses:

- **DSCI 551: Descriptive Statistics and Probability for Data Science**, for basic statistical concepts and familiarity with the mathematical notation.
- **DSCI 552: Statistical Inference and Computation I**, for statistical inference concepts with a frequentist approach.

- **DSCI 561: Regression I**, since the topics of this course follow the same thread.
- **DSCI 531: Data Visualization I**, for plotting tools using the package `ggplot2`.

Dataset References

These are the papers from which each dataset used in the lectures comes from. If you are interested in knowing more about them, you can obtain a PDF copy of each paper with your CWL account via the UBC library:

- Brockmann, H.J. (1996). Satellite Male Groups in Horseshoe Crabs, Limulus polyphemus. *Ethology*, 102: 1-21.
- Deb, P. and Trivedi, P. (1997). Demand for medical care by the elderly: a finite mixture approach. *Journal of Applied Econometrics*, 12(3), 313-336.
- Grunfeld Y. (1958). The determinants of corporate investment. Ph.D. thesis, Department of Economics, University of Chicago.
- Harrison, D. and Rubinfeld, D.L. (1978). Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5, 81–102.
- Mangasarian, O. L., Street, W. N., and Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, 43(4), 570-577.
- Wolberg, W. H. and Mangasarian, O. L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the National Academy of Sciences of the United States of America*, 87(23), 9193–9196.
- Wood, P. (1967). Algebraic Model of the Lactation Curve in Cattle. *Nature*, 216, 164–165.

Policies

See the general MDS policies.

Attribution

The course is built upon previous years' materials developed by previous instructors.

License