The Regression Cookbook

Now with Machine Learning and Stats Flavours!

G. Alexi Rodríguez-Arelis

2024-07-08

This book aims to set a common ground between Machine Learning and Statistics regarding regression techniques, using `Python` and `R`, under two perspectives: **inference** and **prediction**.

Table of contents

Preface

Let the regression cooking begin!

Throughout my journey as a postdoctoral fellow in the Master of Data Science (MDS) at the University of British Columbia, I became aware of the fascinating overlap between **Machine Learning** and **Statistics**. Many **Data Science** students usually come across common Machine Learning/Statistics concepts or ideas that might only differ in names. For instance, simple terms such as weights in **supervised learning** (and their equivalent statistical counterpart as regression coefficients) might be misleading for students starting their Data Science formation. On the other hand, from an instructor's perspective in a Data Science program that subsets its courses in Machine Learning in `Python` and Statistics in `R`, regression courses in `R` also demand the inclusion of `Python`-related packages as alternative tools. In my MDS teaching experience, this is especially critical for students whose career plans lean towards industry where `Python` is more heavily used.

As a Data Science educator, I view this field as a substantial **synergy** between Machine Learning and Statistics. Nevertheless, many gaps between both disciplines still need to be addressed. Thus, closing these critical gaps is imperative in a domain with accelerated growth, such as Data Science. In this regard, the MDS Stat-ML dictionary inspired me to write this textbook. It basically consists of **common ground** between **foundational supervised learning models** from Machine Learning and **regression models** commonly used in Statistics. I strive to explore **linear modelling approaches** as a primary step while highlighting different terminology found in both fields. Furthermore, this discussion is more comprehensive than a simple conceptual exploration. Hence, the second step is **hands-on practice** via the corresponding `Python` packages for Machine Learning and `R` for Statistics.

> Fun Fact!
>
> While thinking about possible names for this work, I was planning to name it "*Machine Learning and Statistics: A Common Ground.*" Nevertheless, it was quite plain and boring! That said, this whole textbook idea sounded analogous to a cookbook, given its heavily applied focus.
> **Hence, the cookbook name idea**!

Audience and Scope

This book mainly focuses on regression analysis and its supervised learning counterpart. Thus, it is not introductory Statistics and Machine Learning material. Instead, the following topics are suggested as prerequisites:

- **Mutivariable Differential Calculus and Linear Algebra.** Certain sections of each chapter pertain to modelling estimation. Therefore, topics such as partial derivatives and matrix algebra are a great asset. You can find helpful learning resources on the MDS webpage.
- **Basic `Python` programming.** When necessary, `Python` {pandas} library will be used to perform data wrangling. The MDS course DSCI 511 (Programming for Data Science) is an ideal example of this prerequisite.

- **Basic `R` programming.** Knowledge of data wrangling and plotting through `R` {tidyverse} is recommended for hands-on practice via the examples provided in each one of the chapters of this book. The MDS courses DSCI 523 (Programming for Data Manipulation) and DSCI 531 (Data Visualization I) are ideal examples of this prerequisite.

- **Foundations of probability and basic distributional knowledge.** The reader should be familiar with elemental discrete and continuous distributions since they are a vital component of any given regression or supervised learning model. The MDS course DSCI 551 (Descriptive Statistics and Probability for Data Science) is an ideal example of this prerequisite.
- **Foundations of frequentist statistical inference.** One of the Data Science paradigms to be covered in this book is statistical inference, i.e., identifying relationships between different variables in a given population or system of interest via a sampled dataset. I only aim to cover a frequentist approach using inferential tools such as parameter estimation, hypothesis testing, and confidence intervals. The MDS course DSCI 552 (Statistical Inference and Computation I) is an ideal example of this prerequisite.
- **Foundations of supervised learning.** The second Data Science paradigm to be covered pertains to prediction, which is core in Machine Learning. The reader should be familiar with basic terminology, such as training and testing data, overfitting, underfitting, cross-validation, etc. The MDS course DSCI 571 (Machine Learning I) provides these foundations.
- **Foundations of feature and model selection.** This prerequisite also relates to Machine Learning and its corresponding prediction paradigm. Basic knowledge of prediction accuracy and variable selection tools is recommended. The MDS course DSCI 573 (Feature and Model Selection) is an ideal example of this prerequisite.

Even though it is recommended that you check the above general topics, especially the statistical ones, I will initially provide a fair review of probability and frequentist inference so we can explain key concepts in data modelling in the context of regression analysis.

How this Book is Structured

This book is structured in three big pillars:

1. The use of an ordered *Data Science workflow*,
2. the correct use of an *appropriate regression model* for supervised learning based on the outcome of interest, and
3. a constant common ground in terminology between Statistics and Machine Learning via a *dictionary*. Each one of these three pillars is heavily connected since the same Data Science workflow is applied in each one of the regression models, which aims to help the reader in their learning (i.e., we would be able to know what exact stage to expect in our data analysis regardless of the regression model we are being exposed to).

On the other hand, the dictionary will allow us to clarify any potential confusion between the two disciplines, Statistics and Machine Learning, in differences in terminology within supervised learning. As a side note, throughout the book, all statistical terms will be highlighted in **purple** whereas the common Machine Learning terms will be highlighted in **blue**. This colour scheme strives to combine this terminology so we can switch from one field to another in an easier way. With practice and time, the reader should be expected to jump back and forth when using these concepts.

Data Science Workflow

A crucial aspect of the practice of supervised learning is the need for a systematic Data Science workflow that will allow us to solve our respective inquiries in a transparent and reproducible way. Hence, **?@sec-intro** begins with the general definition of supervised learning and its two critical paradigms: *inference* and *prediction*. Moreover, its workflow is fully explained across seven sequential phases:

i. **Design**.
ii. **Data collection**.
iii. **Exploratory data analysis**.
iv. **Modelling**.
v. **Estimation**.
vi. **Results**.
vii. **Storytelling**.

Now, suppose we do not follow a predefined workflow in practice. In that case, we might be at risk of incorrectly addressing these inquiries, translating into meaningless results outside the context of the problem we aim to solve. This is why the formation of a Data Scientist must stress this workflow from the very introductory stages.

On the other hand, even though this book has prerequisites related to the basics of probability via different distributions and the fundamentals of frequentist statistical inference, **?@sec-intro** also provides a handy summary of all these topics. We will check essential concepts such as *random variables*, *sampling*, *parameter estimation via maximum likelihood*, the process of *hypothesis testing*, delivery of *confidence intervals*, and how to obtain *outcome predictions*. Note these concepts are heavily related to the Data Science workflow depicted in **?@sec-ds-workflow** across its different phases.

Mindmap of Regression Analysis

Having defined the necessary statistical aspects to execute a proper supervised learning analysis, either *inferential* or *predictive* across its seven sequential phases, we must dig into the different approaches we might encounter in practice as regression models. The nature of our outcome of interest will dictate any given modelling approach to apply, depicted as clouds in Figure 1. Note these regression models can be split into two sets depending on whether the outcome of interest is *continuous* or *discrete*. Therefore, under a probabilistic view, identifying the nature of a given random variable is crucial in regression analysis.

As we can see in the clouds of Figure 1, there are 13 regression models: 8 belonging to discrete outcomes and 5 to continuous outcomes. Each of these models is contained in a chapter of this book, beginning with the most basic regression tool known as ordinary least-squares in **?@sec-ols**. We must clarify that the current statistical literature is not restricted to these 13 regression models. The field of regression analysis is vast, and one might encounter more complex models to target certain specific inquiries. Nonetheless, I consider these models the fundamental regression approaches that any Data Scientist must be familiar with in everyday practice.

Dictionary of Terminology

Finally, following the structure of the MDS Stat-ML dictionary, **?@sec-dictionary** contains a dictionary that follows the same format which contrasts the terminology a Data Scientist can find in Machine Learning and Statistics (indicated as **ML/Stats** throughout the book) when performing supervised learning in practice. Note that terms will be coloured depending on their **Machine Learning** or **statistical** nature.

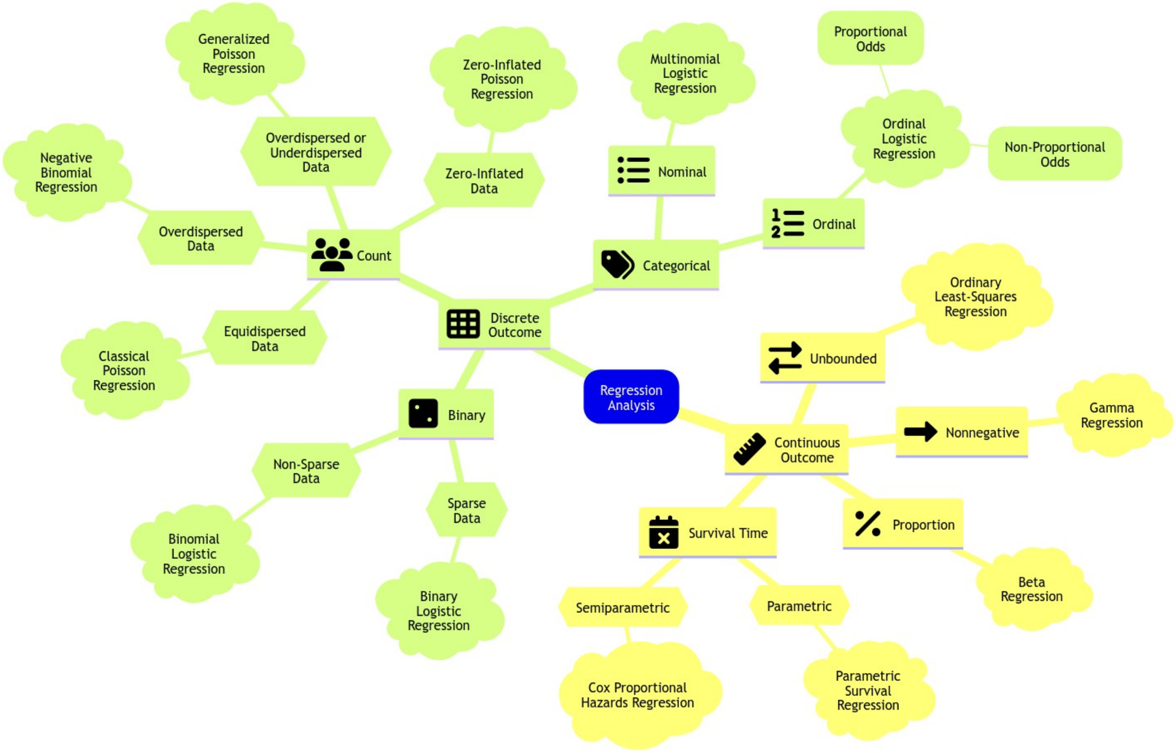
Regression Analysis
Discrete Outcome
Count
Overdispersed or Underdispersed Data
Generalized Poisson Regression
Overdispersed Data
Negative Binomial Regression
Equidispersed Data
Classical Poisson Regression
Zero-Inflated Data
Zero-Inflated Poisson Regression
Categorical
Nominal
Multinomial Logistic Regression
Ordinal
Ordinal Logistic Regression
Proportional Odds
Non-Proportional Odds
Binary
Non-Sparse Data
Binomial Logistic Regression
Sparse Data
Binary Logistic Regression
Continuous Outcome
Unbounded
Ordinary Least-Squares Regression
Nonnegative
Gamma Regression
Proportion
Beta Regression
Survival Time
Semiparametric
Cox Proportional Hazards Regression
Parametric
Parametric Survival Regression

Figure 1: Regression analysis mindmap depicting all modelling techniques to be explored in this book. These techniques are split into two big sets: *continuous* and *discrete* outcomes.

Heads-up!

Readers might be surprised that certain concepts refer to the same idea but with different terminology. Contrarily, in some other cases, the same terminology might refer to different concepts!

1 Introduction

This is a book created from markdown and executable code.

See Knuth (1984) for additional discussion of literate programming.

```
1 + 1
```

```
[1] 2
```

2 Summary

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```

References

Knuth, Donald E. 1984. “Literate Programming.” *Comput. J.* 27 (2): 97–111. https://doi.org/10.1093/comjnl/27.2.97.