The Regression Cookbook

Now with Machine Learning and Stats Flavours!

G. Alexi Rodríguez-Arelis

2024-08-08

This book aims to set a common ground between machine learning and statistics regarding linear 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 data science teaching experience, this is especially critical for students whose career plan leans 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**[1], given its heavily applied focus.

Hence, the cookbook name idea!

[1]Special thanks to *Jonathan Graves*, who mentioned this word in our very first chat back when I was conceptualizing this textbook.

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. Also, some coding background on `R` (R Core Team 2024) and/or `Python` (Van Rossum and Drake 2009) is recommended. That said, the following topics are suggested as **fundamental reviews**:

- **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} (The Pandas Development Team 2024) library will be used to perform *data wrangling*. The MDS course DSCI 511 (Programming for Data Science) is an ideal example of a quick review.

Figure 1: Image by *Lubos Houska* via *Pixabay*.

- **Basic `R` programming.** Knowledge of *data wrangling and plotting* through `R` {tidyverse} (Wickham et al. 2019) is recommended for hands-on practice via the cases 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 a quick review.
- **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 a quick review.
- **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 a quick review.
- **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 a quick review.

A further remark on probability and statistical inference

In case the reader is not 100% familiar with probabilistic and inferential topics, as discussed above, I will provide a fundamental refresher in Chapter 1 with crucial points that are needed to follow along the **statistical way** each one of the chapters is delivered (more specifically for **modelling estimation/training** matters!).

Furthermore, this refresher will be integrated into **the three big pillars** that will be fully expanded in this book: a **data science workflow**, the right **workflow flavour** (**inferential** or **predictive**), and a **regression toolbox**.

1 Getting Ready for Regression Cooking!

It is time to prepare for the different regression techniques we will use in each of the subsequent chapters of this book. That said, there is a strong message I want to convey across all this work:

> **Different modelling estimation techniques in regression analysis can be smoothly grasped when we develop a fair probabilistic and inferential intuition on our populations or systems of interest.**

The above statement has a key statistical foundation on how data is generated and can be modelled via different regression approaches. More details on the concepts and ideas associated with this foundation will be delivered in Section 1.2.

Figure 1.1: Image by *Lucas Israel* via *Pixabay*.

Then, once we have reviewed these statistical concepts and ideas, we will move on to the three big pillars I previously pointed out:

1. The use of an ordered **data science workflow** in Section 1.3,

2. choosing the proper workflow flavour according to either an **inferential** or **predictive** paradigm as shown in Figure 1.7, and
3. the correct use of an **appropriate regression model** based on the response or outcome of interest as shown in the mind map from Section 1.4 (analogous to a **regression toolbox**).

The Rationale Behind the Three Pillars

Each data science-related problem that uses regression analysis might have distinctive characteristics considering inferential (statistics!) or predictive (machine learning!) inquiries. Specific problems would implicate using outcomes (or responses) related to survival times (e.g., the time until one particular equipment of a given brand fails), categories (e.g., a preferred musical genre in the Canadian young population), counts (e.g., how many customers we would expect on a regular Monday morning in a national central bank), etc. Moreover, under this regression context, our analyses would be expanded to explore and assess how our outcome of interest is related to a further set of variables (the so-called features!). For instance, following up with the categorical outcome of the preferred musical genre in the Canadian young population, we might analyze how particular age groups prefer certain genres over others or even how preferred genres compare each other across different Canadian provinces in this young population. **The sky is the limit here!**

Therefore, we might be tempted to say that each regression problem should have its own workflow, given that the regression model to use would implicate particular analysis phases. **However, it turns out that is not the case to a certain extent**, and we have a regression workflow in Figure 1.7 to support this bold statement as a proof of concept for thirteen different regression models (i.e, thirteen subsequent chapters in this book). The workflow aims to homogenize our data analyses and make our modelling process more transparent and smoother. We can deliver exactly concluding insights as data storytelling while addressing our initial main inquires. Of course, when depicting the workflow as a flowchart, there will be decision points that will turn it into **inferential** or **predictive** (the second pillar). Finally, where does the third pillar come into play in this workflow? This pillar is contained in the **data modelling stage**, where the mind map from Figure 1.16 will come in handy.

Now, before delving into probability and frequentist statistical inference, let us establish a convention on the use of admonitions beginning Section 1.2 in this textbook:

! Important 1: Definition

A formal statistical and/or machine learning definition. This admonition aims to untangle the significant amount of jargon and concepts that both fields have. Furthermore,

alternative terminology will be brought up when necessary to indicate the same definition across both fields.

Note 1: Heads-up!

An idea (or ideas!) of key relevance for any given modelling approach, specific workflow stage or data science-related terminology. This admonition also extends to crucial statistical or machine learning topics that the reader would be interested in exploring more in-depth.

Tip 1: Tip

An idea (or ideas!) that might be slightly out of the scope of the topic any specific section is discussing. Still, I will provide significant insights on the matter along with further literature references to look for.

The core idea of the above admonition arrangement is to allow the reader to discern between ideas or concepts that are key to grasp from those whose understanding might not be highly essential (but still interesting to check out in external references!).

1.1 The ML-Stats Dictionary

The above admonition in Important 1 for a **definition** will pave the way to a complimentary aspect of this textbook that I have had in mind since I started teaching statistics (and, more especially, regression analysis) in a data science context. Machine learning and statistics usually overlap across many subjects, and regression modelling is no exception. Topics we teach in an utterly regression-based course, under a purely statistical framework, also appear in machine learning-based courses such as fundamental supervised learning, but often with different terminology. On this basis, the Master of Data Science (MDS) program at the University of British Columbia (UBC) provides the MDS Stat-ML dictionary (Gelbart 2017) under the following premises:

> *This document is intended to help students navigate the large amount of jargon, terminology, and acronyms encountered in the MDS program and beyond.*
>
> *This section covers terms that have different meanings in different contexts, specifically statistics vs. machine learning (ML).*

Indeed, both disciplines have a tremendous amount of jargon and terminology. Furthermore, as I previously emphasized in the **Preface**, machine learning and statistics construct a **substantial synergy** that is reflected in data science. Even with this, people in both fields

could encounter miscommunication issues when working together. This should not happen if we build solid bridges between both disciplines. Hence, a comprehensive **ML-Stats dictionary** (*ML* stands for *Machine Learning*) is imperative, and this textbook offers a perfect opportunity to build this resource. Primarily, this dictionary clarifies any potential confusion between statistics and machine learning regarding terminology within supervised learning and regression analysis contexts.

> **Note 2: Heads-up on terminology highlights!**
>
> Following the spirit of the **ML-Stats dictionary**, throughout the book, all statistical terms will be highlighted in magenta whereas the machine learning terms will be highlighted in orange. 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, we should be able to jump back and forth when using these concepts.

Finally, Appendix A will be the section in this book where the reader can find all those statistical and machine learning-related terms in alphabetical order. Notable terms (either statistical or machine learning-related) will include an admonition identifying which terms (again, either statistical or machine learning-related) are **equivalent** (**or NOT equivalent if that is the case**). Take as an example the statistical term dependent variable:

> In supervised learning, it is the main variable of interest we are trying to **learn** or **predict**, or equivalently, the variable we are trying **explain** in a statistical inference framework.

Then, the above definition will be followed by this admonition:

> **Equivalent to:**
>
> Response, outcome, output or target.

Note we have identified four equivalent terms for the term dependent variable. Furthermore, according to our already defined colour scheme, these terms can be statistical or machine learning-related.

> **Note 3: Heads-up on the use of terminology!**
>
> Throughout this book, I will interchangeably use specific terms when explaining the different regression approaches in each chapter. Whenever confusion arises about using these interchangeable terms, it is highly recommended to consult their definitions and equivalences (or non-equivalences) in Appendix A.

Now, let us proceed to a quick review on probability and statistics in a frequentist framework. This review will be important to understanding the philosophy of modelling parameter

estimation, mainly in relation to statistical inference.

1.2 A Quick Review on Probability and Frequentist Statistical Inference

When I was an undergraduate student and took my very first course in probability and statistics (inference included!) in an industrial engineering context, I used to feel quite overwhelmed by the large amount of jargon and formulas one had to grasp and use regularly for primary engineering fields such as quality control in a manufacturing facility. *Population parameters*, *hypothesis testing*, *tests statistics*, *significance level*, *p-values*, and *confidence intervals* (**do not worry, our statistical/machine learning scheme will come in later in this quick review**) were appearing here and there! And to my frustration, I could never find a statistical connection between all these inferential tools! Instead, I relied on mechanistic procedures when solving assignments or exam problems.

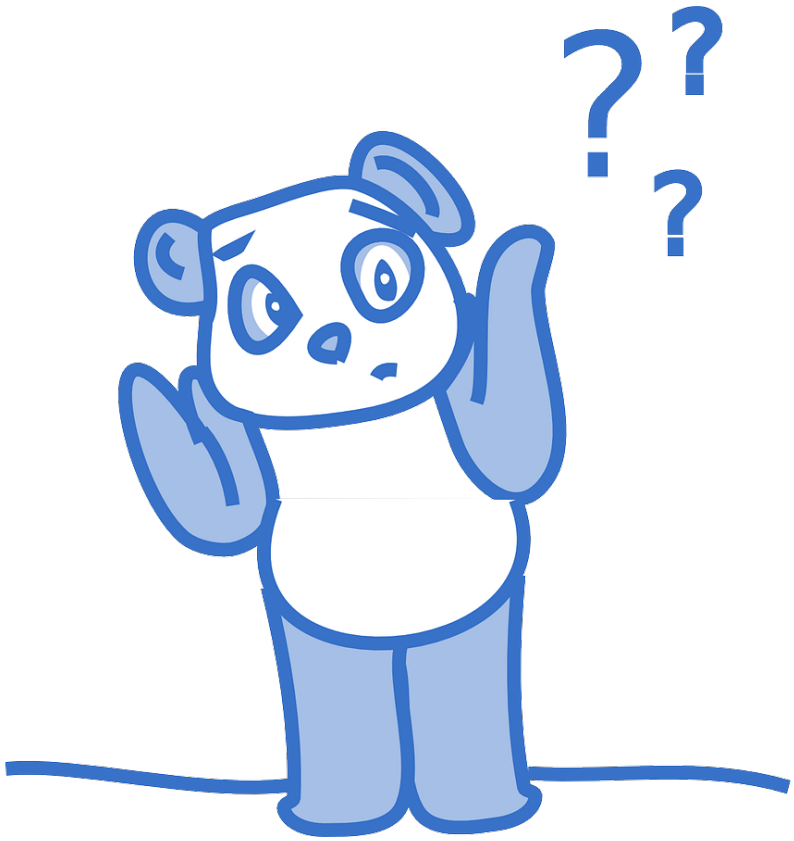

Figure 1.2: Image by *OpenClipart-Vectors* via *Pixabay*.

For instance, when performing hypothesis testing for a two-sample t-test, I never reflected on what the hypotheses were trying to indicate for the corresponding population parameters(s) nor how the test statistic was related to these hypotheses. Moreover, my interpretation of the resulting p-value and/or confidence interval was purely mechanical with the inherent claim:

> *With a significance level $\alpha = 0.05$, we reject (**or fail to reject, if that is the case!**) the null hypothesis given that…*

Honestly, I am not proud of this whole mechanical way of doing statistics now that I reflect on it after many years of practice, teaching, and research. Then, of course, the above situation should not happen when we learn key statistical topics for the very first time as undergraduate students. That is why we will dig into a more intuitive way of viewing probability and its crucial role in statistical inference. This matter will help us deliver more coherent storytelling when presenting our results in practice during any regression analysis to our peers or stakeholders. Note that the role of probability also extends to model training when it comes to supervised learning and not just in regard to statistical inference.

Having said all this, it is time to introduce a statement that I ended up reasoning the very first time I taught hypothesis testing in the introductory statistical inference course in the MDS program at UBC:

> **In statistical inference, everything always boils down to randomness and how we can control it!**

That is quite a bold statement! Nonetheless, once one starts teaching statistical topics to audiences not entirely familiar with the usual field jargon, the idea of randomness always persists across many different tools. And, of course, regression analysis is not an exception at all since it also involves inference on population parameters of interest! This is why I have allocated this section in the textbook to explain core probabilistic and inferential concepts to pave the way to its role in regression analysis.

i Note 4: Heads-up on why we mean as a non-ideal mechanical analysis!

The reader might get a contradictory impression of why the above mechanical way of performing hypothesis testing is viewed as something **not ideal**, whereas the word **cookbook** appears in this book's title. Certainly, the **cookbook** idea refers to some class of recipe to perform data modelling (which is reflected in the workflow from Figure 1.7). Still, there is a critical difference between a non-ideal mechanical way of performing hypothesis testing versus our data modelling workflow.

On the one hand, the non-ideal mechanical way refers to **the use of a tool without understanding the rationale of what this tool stands for**, resulting in vacuous and standard statements that we would not be able to explain any way further, such as the statement I previously indicated:

> *With a significance level $\alpha = 0.05$, we reject (**or fail to reject, if that is the case!**) the null hypothesis given that...*
>
> What if a stakeholder of our analysis asks us in plain words what a significance level means? Why are we phrasing our conclusion on the null hypothesis and not directly on the alternative one? As a data scientist, one should be able to deliver the corresponding explanations of why the whole inference process is yielding that statement without misleading the stakeholder's understanding. **For sure, this also implicates appropriate communication skills that cater to general audiences rather than just statistical.**
>
> On the other hand, as we will elaborate in further detail, the workflow in Figure 1.7 has stages that demand a **thorough and precise understanding of what exactly we are executing in our analysis**. Moving forward from one current stage in the workflow to the next (without a complete understanding of what is going on in the current one) would risk carrying over false insights that might become **faulty data storytelling** of the whole analysis.

Finally, even though this book has suggested reviews related to the basics of probability via different distributions and the fundamentals of frequentist statistical inference as stated in **Audience and Scope**, we will retake essential concepts as follows:

- The role of *random variables* and *probability distributions* and the governance of *population (or system) parameters* (i.e., the so-called Greek letters we usually see in statistical inference and regression analysis). Section 1.2.1 will explore these topics more in detail while connecting them to the subsequent inferential terrain under a *frequentist context*.
- When delving into supervised learning and regression analysis, we might wonder how randomness is incorporated into *model fitting* (i.e., *parameter estimation*). That is quite a fascinating aspect, implemented via a crucial statistical tool known as *maximum likelihood estimation*. This tool is heavily related to the concept of *loss function* in supervised learning. Section 1.2.2 will explore these matters in more detail and how the concept of *random sample* is connected to this estimation tool.
- Section 1.2.3 will explore the basics of *hypothesis testing* and its intrinsic components such as *null* and *alternative hypotheses*, *type I* and *type II* errors, *test statistic*, *standard error*, *p-value*, and *confidence interval*.
- Finally, Section 1.2.4 will briefly discuss the connections between supervised learning and regression analysis regarding terminology.

Without further ado, let us start with reviewing core concepts in probability via quite a tasty example.

1.2.1 Basics of Probability

In terms of regression analysis and its supervised learning counterpart (either on an **inferential** or **predictive** framework), probability can be viewed as the solid foundation on which more complex tools, including estimation and hypothesis testing, are built upon. Under this foundation, our data is coming from a given population or system of interest. Moreover, the population or system is assumed to be governed by parameters which, as data scientists or researchers, they are of their best interest to study. That said, the terms population and parameter will pave the way to our first statistical definitions.

> ❗ Important 2: Definition of population
>
> It is a **whole collection of individuals or items** that share **distinctive attributes**. As data scientists or researchers, we are interested in studying these attributes, which we assume are **governed** by parameters. In practice, we must be **as precise as possible** when defining our given population such that we would frame our entire data modelling process since its very early stages. Examples of a population could be the following:
>
> - *Children between the ages of 5 and 10 years old in states of the American West Coast.*
> - *Customers of musical vinyl records in the Canadian provinces of British Columbia and Alberta.*
> - *Avocado trees grown in the Mexican state of Michoacán.*
> - *Adult giant pandas in the Southwestern Chinese province of Sichuan.*
> - *Mature açaí palm trees from the Brazilian Amazonian jungle.*

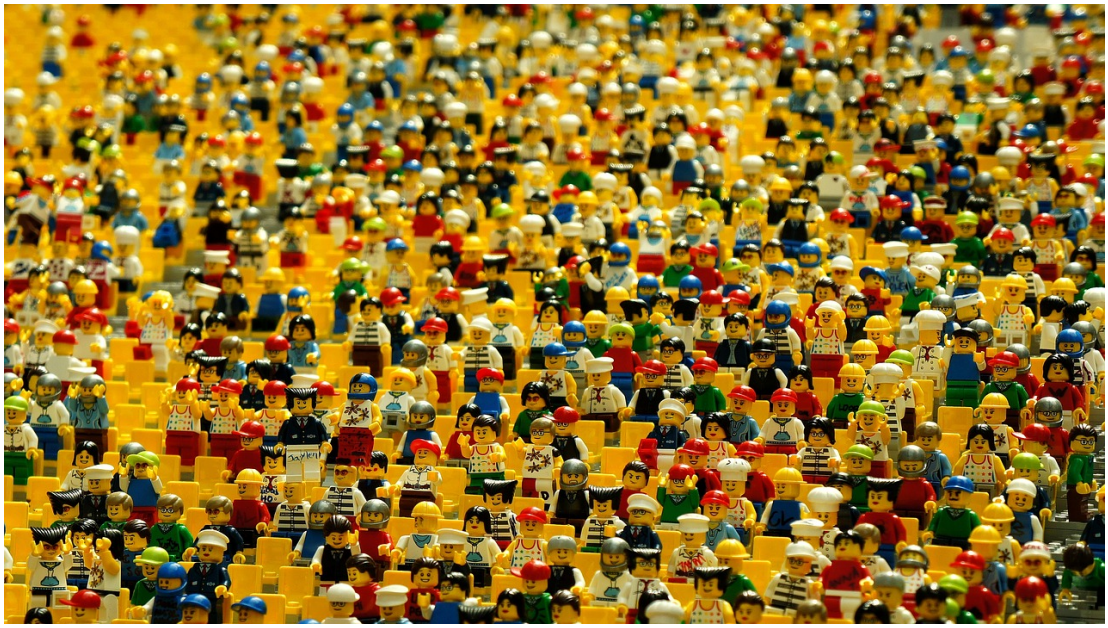

Figure 1.3: Image by *Eak K.* via *Pixabay*.

Note that the term population could be exchanged for the term **system**, given that certain contexts do not particularly refer to individuals or items. Instead, these contexts could refer to **processes** whose attributes are also governed by parameters. Examples of a **system** could be the following:

- *The production of cellular phones from a given model in a set of manufacturing facilities.*
- *The sale process in the Vancouver franchises of a well-known ice cream parlour.*
- *The transit cycle during rush hours on weekdays in the twelve lines of Mexico City's subway.*

! Important 3: Definition of parameter

It is a characteristic (**numerical** or even **non-numerical**, such as a **distinctive category**) that **summarizes** the state of our population or **system** of interest. Examples of a population parameter can be described as follows:

- *The average weight of children between the ages of 5 and 10 years old in states of the American west coast.*
- *The variability in the height of the mature açaí palm trees from the Brazilian Amazonian jungle.*

- *The proportion of defective items in the production of cellular phones in a set of manufacturing facilities.*
- *The average customer waiting time to get their order in the Vancouver franchises of a well-known ice cream parlour.*
- *The most favourite pizza topping of vegetarian adults between the ages of 30 and 40 years old in Edmonton.*

Figure 1.4: Image by *meineresterampe* via *Pixabay*.

Note the **standard mathematical notation** for a population parameters are **Greek letters**. Moreover, in practice, these population parameter(s) of interest will be **unknown** to the data scientist or researcher. Instead, they would use formal statistical inference to **estimate** them.

The parameter definition in Important 3 points out a crucial fact in investigating any given population or system:

Our parameter(s) of interest are usually *unknown*!

Given this fact, it would be pretty unfortunate and inconvenient if we eventually wanted to discover any significant insights about the population or system. Therefore, let us proceed to our so-called tasty example so we can dive into the need for statistical inference and why probability is our perfect ally in this parameter quest.

Imagine you are the owner of a large fleet of ice cream carts, around 900 to be exact. These ice

cream carts operate across different parks in the following Canadian cities: *Vancouver*, *Victoria*, *Edmonton*, *Calgary*, *Winnipeg*, *Ottawa*, *Toronto*, and *Montréal*. In the past, to optimize operational costs, you decided to limit ice cream cones to only two items: *vanilla* and *chocolate* flavours, as in Figure 1.5.

Figure 1.5: The two flavours of the ice cream cone you sell across all your ice cream carts: *vanilla* and *chocolate*. Image by *tomekwalecki* via *Pixabay*.

Now, let us direct this whole case onto a more statistical and probabilistic field; suppose you have a well-defined overall population of interest for those above eight Canadian cities: **children between 4 and 11 years old attending these parks during the Summer weekends**. Of course, Summer time is coming this year, and you would like to know **which ice cream cone flavour is the favourite one** for this population (*and by how much!*). As a business owner, investigating ice cream flavour preferences would allow you to plan Summer

restocks more carefully with your corresponding suppliers. Therefore, it would be essential to start collecting consumer data so the company can tackle this **demand query**.

Also, suppose there is a second query. For the sake of our case, we will call it a **time query**. As a critical component of demand planning, besides estimating which cone flavour is the most preferred one (*and by how much!*) for the above population of interest, the operations area is currently requiring a realistic estimation of **the average waiting time from one customer to the next one in any given cart during Summer weekends**. This average waiting time would allow the operations team to plan carefully how much stock each cart should have so there will not be any waste or shortage.

Figure 1.6: Image by *Icons8 Team* via *Unsplash*.

Note that the nature of the aforementioned **time query** is more related to a larger population. Therefore, we can define it as **all our ice cream customers during the Summer weekends**. Furthermore, this second definition would limit this query to our corresponding general ice cream customers, given the requirements of our operations team, and **not** all the children between 4 and 11 years old attending the parks during Summer weekends. Consequently, it is crucial to note that the nature of our queries will dictate how we define our population and our subsequent data modelling and statistical inference.

Summer time represents the most profitable season from a business perspective, thus solving these above two queries is a significant priority for our company. Hence, you decide to organize

a meeting with your eight general managers (one per Canadian city). Finally, during the meeting with the general managers, it was decided to do the following:

1. For the **demand query**, a comprehensive market study will be run on the population of interest across the eight Canadian cities right before next Summer; suppose we are currently in Spring.
2. For the **time query**, since the operations team has not previously recorded any historical data, **ALL** vendor staff from 900 carts will start collecting data on **the waiting time in seconds** between each customer this upcoming Summer.

Surprisingly, when discussing study requirements for the marketing firm who would be in charge of it for the **demand query**, *Vancouver's general manager* dares to state the following:

> *Since we're already planning to collect consumer data on these cities, let's mimic a census-type study to ensure we can have the **MOST PRECISE** results on their preferences.*

On the other hand, when agreeing on the specific operations protocol to start recording waiting times for all the 900 vending carts this upcoming Summer, *Ottawa's general manager* provides a comment for further statistical food for thought:

> *The operations protocol for recording waiting times in the 900 vending carts looks too cumbersome to implement straightforwardly this upcoming Summer. Why don't we select **A SMALLER GROUP** of ice cream carts across the eight cities to have a more efficient process implementation that would allow us to optimize operational costs?*

Bingo! *Ottawa's general manager* just nailed the probabilistic way of making inference on our population parameter of interest for the **time query**. Indeed, their comment was primarily framed from a business perspective of optimizing operational costs. Still, this fact does not take away a crucial insight on which statistical inference is built: a random sample (as in Important 6). As for *Vancouver's general manager*, ironically, their statement is **NOT PRECISE** at all! Mimicking a census-type study might not be the most optimal decision for the **demand query** given the time constraint and the potential size of its target population.

> **Realistically, there is no cheap and efficient and cheap way to conduct a census-type study for any of the two queries!**

We can state that probability is viewed as the language to decode random phenomena that occur in any given population or system of interest. In our example, we have two random phenomena:

1. For the **demand query**, a phenomenon can be represented by the preferred ice cream cone flavour of **any randomly selected child between 4 and 11 years old attending the parks of the above eight Canadian cities during the Summer weekends**.
2. Regarding the **time query**, a phenomenon of this kind can be represented by **any randomly recorded waiting time between two customers during a Summer weekend in any of the above eight Canadian cities**.

Hence, let us finally define what we mean by probability along with the inherent concept of sample space.

! Important 4: Definition of probability

Let A be an event of interest in a random phenomenon, in a population or **system** of interest, whose all possible outcomes belong to a given sample space S. Generally, the probability for this event A happening can be mathematically depicted as $P(A)$. Moreover, **suppose we observe the random phenomenon n times** such as we were running some class of experiment, then $P(A)$ is defined as the following ratio:

$$P(A) = \frac{\text{Number of times event } A \text{ is observed}}{n}, \tag{1.1}$$

as the n times we observe the random phenomenon goes to infinity.
Equation 1.1 will always put $P(A)$ in the following numerical range:

$$0 \leq P(A) \leq 1.$$

! Important 5: Definition of sample space

Let A be an event of interest in a random phenomenon in a population or **system** of interest. The sample space S of event A denotes the set of all the possible **random outcomes** we might encounter every time we randomly observe A such as we were running some class of experiment.

Note each of these outcomes has a determined probability associated with them. If we add up all these probabilities, the probability of the sample S will be one, i.e.,

$$P(S) = 1. \tag{1.2}$$

Equation 1.3

$$P(X = x \mid \pi) = \pi^x (1 - \pi)^{1-x} \quad \text{for} \quad x = 0, 1. \tag{1.3}$$

❗ Important 6: Definition of random sample

1.2.2 What is Maximum Likelihood Estimation?

1.2.3 Basics of Frequentist Statistical Inference

1.2.4 Supervised Learning and Regression Analysis

1.3 The Data Science Workflow

It is time to review the so-called data science workflow. Each one of these three pillars is heavily connected since a general Data Science workflow is applied in each one of these regression models, which aims to help in our 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). Therefore, a crucial aspect of the practice of Regression Analysis is the need for this systematic Data Science workflow that will allow us to solve our respective inquiries in a reproducible way. Figure 1.7 shows this workflow which has the following general stages (I briefly define each one of them; note a broader delivery will be done in subsequent subsections):

1. **Study design:**
2. **Data collection and wrangling:**
3. **Exploratory data analysis:**
4. **Data modelling:**
5. **Estimation:**
6. **Goodness of fit:**
7. **Results:**
8. **Storytelling**

ℹ What if there is no formal structure in our regression analysis?

Since very early learning stages in data analysis, it is crucial tp
Now, suppose we do not follow a predefined workflow in practice. In that case, we might be at stake in incorrectly addressing our 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 learning stages.

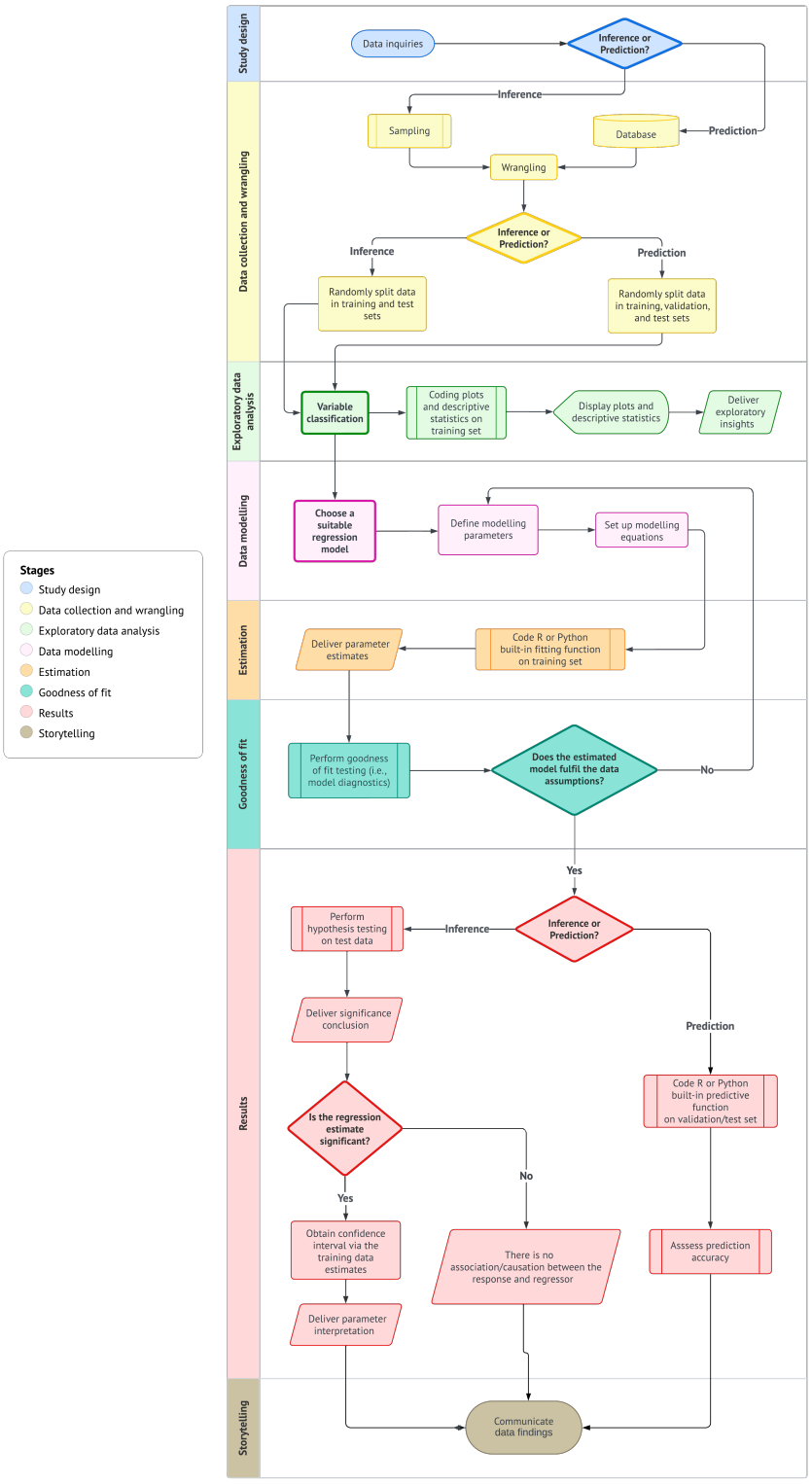

Figure 1.7: Data science workflow for *inferential* and *predictive* inquiries in regression analysis and supervised learning, respectively.

1.3.1 Study Design

The first stage of this workflow is heavily related to the *main statistical inquiries* we aim to address throughout the whole data analysis process. As a data scientist, it is your task to primarily translate these inquiries from the stakeholders of the problem as *inferential* or *predictive*. Roughly speaking, this primary classification can be explained as follows:

- **Inferential.** The main objective is to untangle relationships of *association* or *causation* between the regressors (i.e., explanatory variables) and the corresponding response in the context of the problem of interest. Firstly, we would assess whether there is a statistical relationship between them. Then, if significant, we would quantify by how much.
- **Predictive.** The main objective is to deliver response predictions on further observations of regressors, having estimated a given model via a current training dataset. Unlike inferential inquiries, assessing a statistically significant association or causation between our variables of interest is not a primary objective but *accurate predictions*. **This is one of the fundamental paradigms of machine learning.**

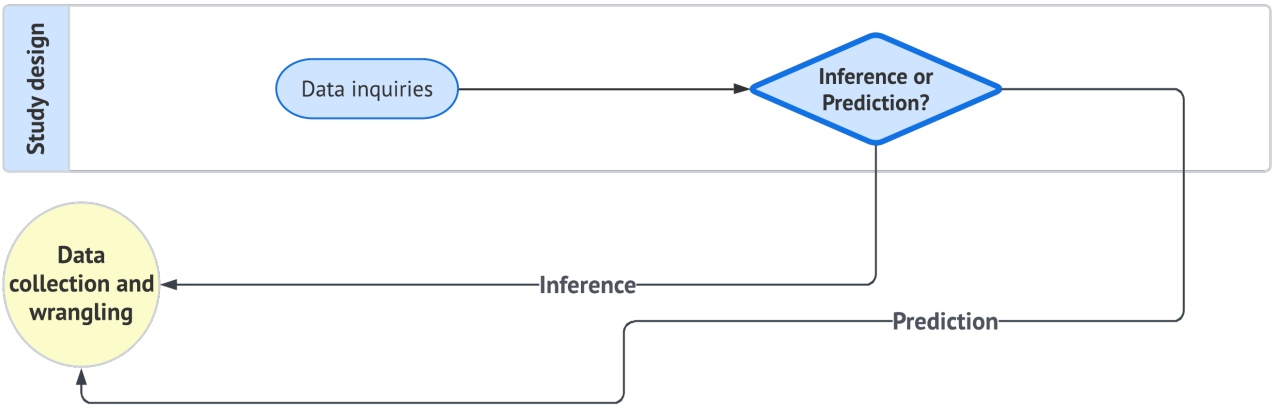

Figure 1.8: *Study design* stage from the data science workflow in Figure 1.7. This stage is directly followed by *data collection and wrangling*.

1.3.2 Data Collection and Wrangling

Once we have defined our main statistical inquiries, it is time to collect our data. Note we have to be careful about the way we collect this data since it might have a particular impact on the quality of our statistical practice:

- Regarding inferential inquiries, recall we are approaching populations or systems of interest governed by *unknown and fixed distributional parameters*. Thus, via sampled data,

we aim to *estimate* these distributional parameters. This is why **a proper sampling method** on this population or system of interest is critical to obtaining representative data for *appropriate hypothesis testing*.

Tip 2: A Quick Debrief on Sampling!

Sampling topics are out of the scope of this book. Nevertheless, we still need to stress that a proper sampling method is also key in inferential inquiries to assess association and/or causation between the regressors and your response of interest. That said, depending on the context of the problem, we could apply either one of the following methods of sampling:

- **Simple random sampling**.
- **Systematic sampling**.
- **Stratified sampling**.
- **Clustered sampling**.
- Etc.

As in the case of Regression Analysis, statistical sampling is a vast field, and we could spend a whole course on it. If you are more interested in these topics, *Sampling: design and analysis* by Lohr offers great foundations.

- In practice, regarding predictive inquiries, we would likely have to deal with databases given that our trained models will not be used to make inference and parameter interpretations.

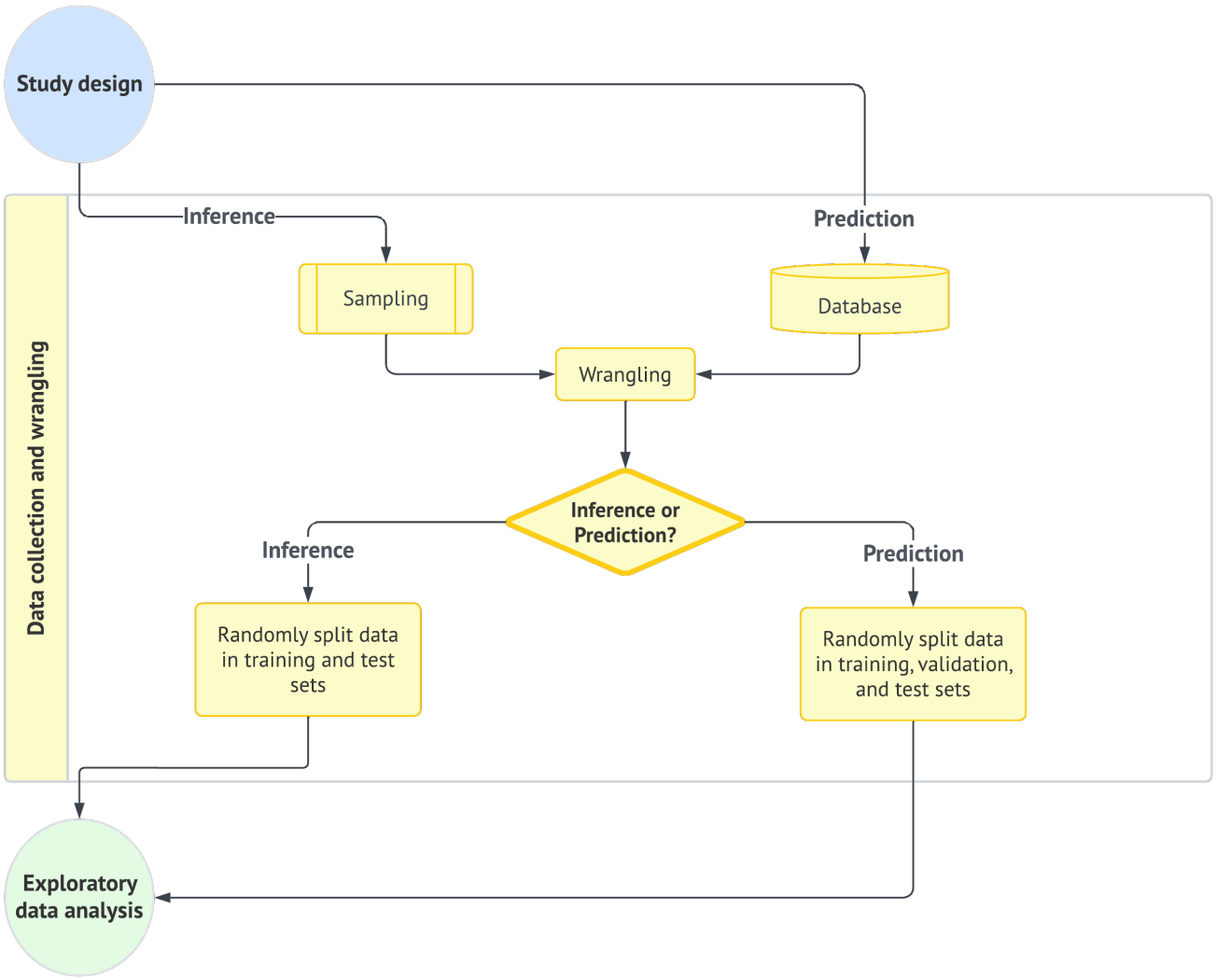

Figure 1.9: *Data collection and wrangling* stage from the data science workflow in Figure 1.7. This stage is directly followed by *exploratory data analysis* and preceded by *study design*.

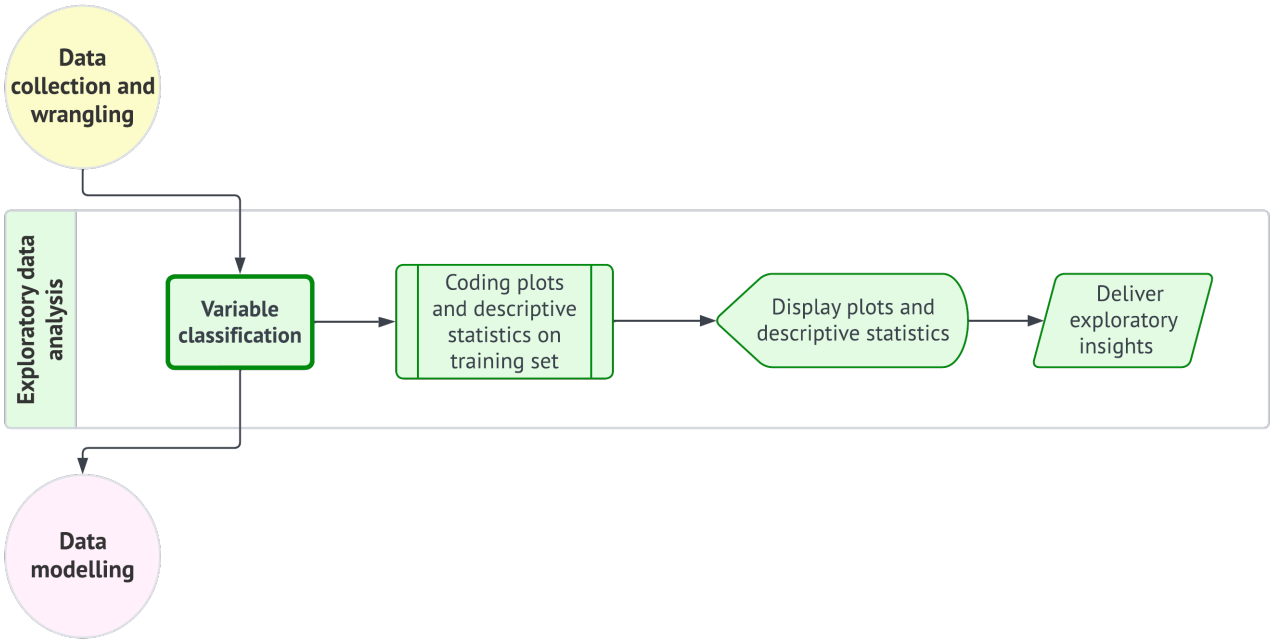

Figure 1.10: *Exploratory data analysis* stage from the data science workflow in Figure 1.7. This stage is directly followed by *data modelling* and preceded by *data collection and wrangling*.

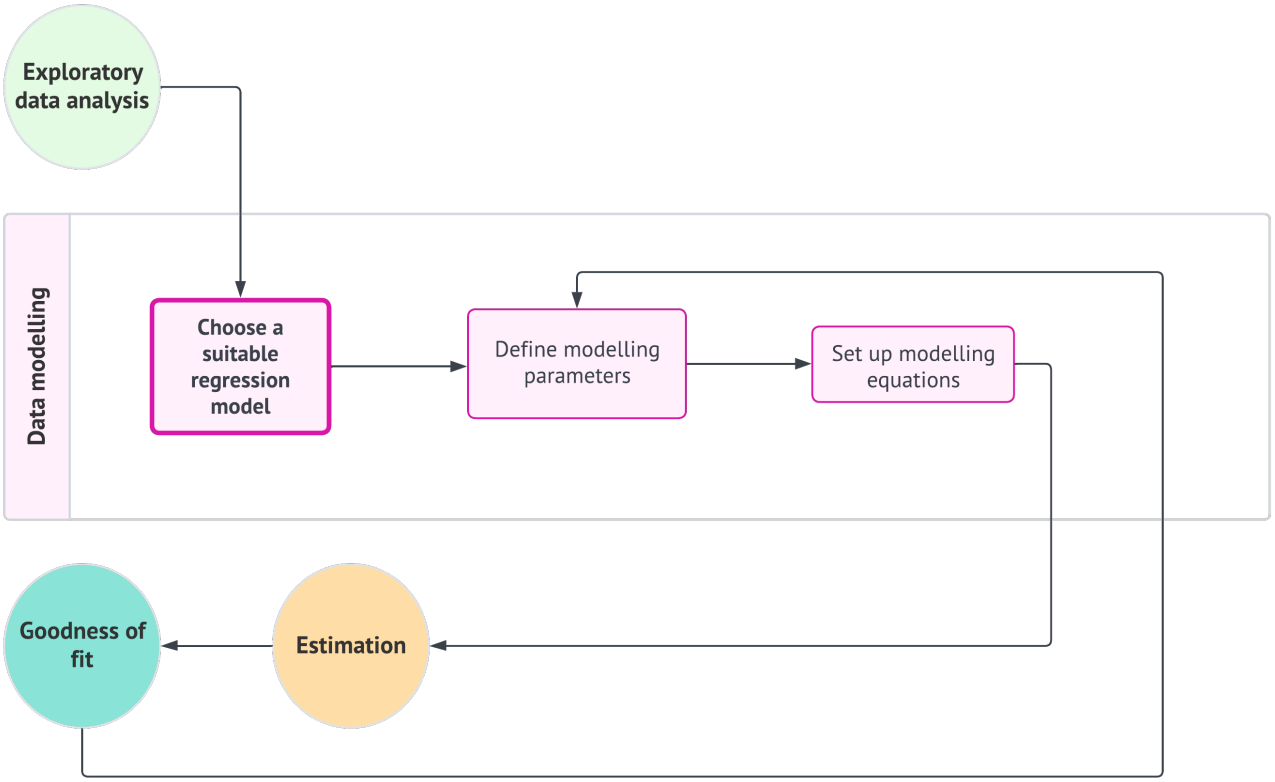

Figure 1.11: *Data modelling* stage from the data science workflow in Figure 1.7. This stage is directly preceded by *exploratory data analysis*. On the other hand, it is directly followed by *estimation* but indirectly with *goodness of fit*. If necessary, the *goodness of fit* stage could retake the process to *data modelling*.

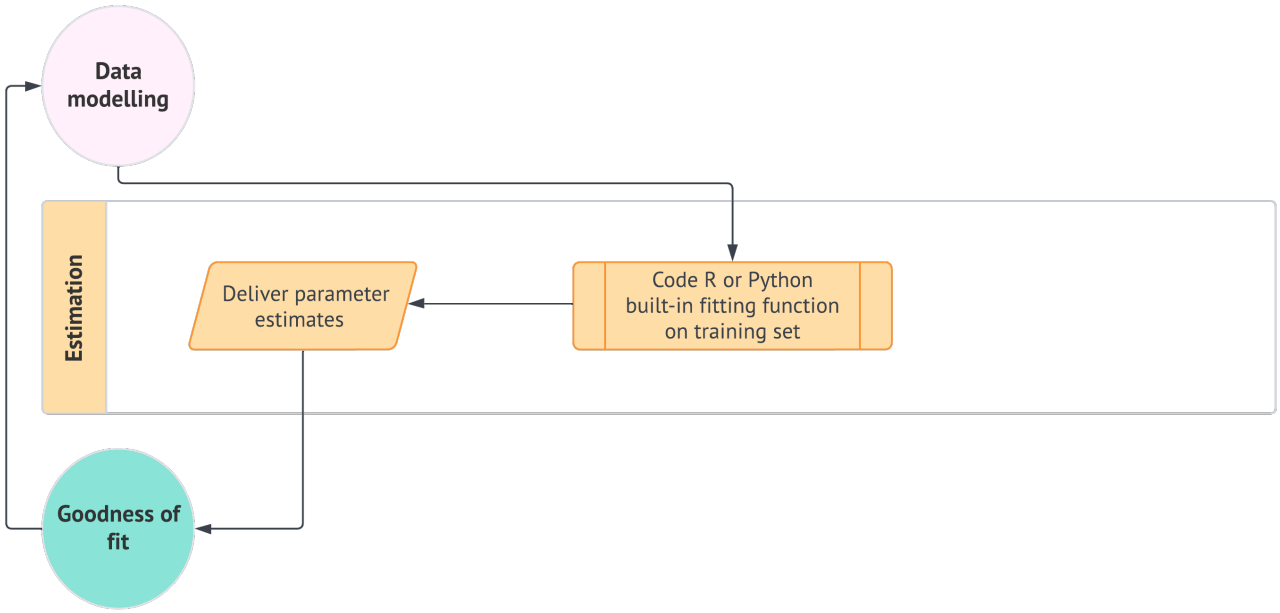

Figure 1.12: *Estimation* stage from the data science workflow in Figure 1.7. This stage is directly preceded by *data modelling* and followed by *goodness of fit*. If necessary, the *goodness of fit* stage could retake the process to *data modelling* and then to *estimation*.

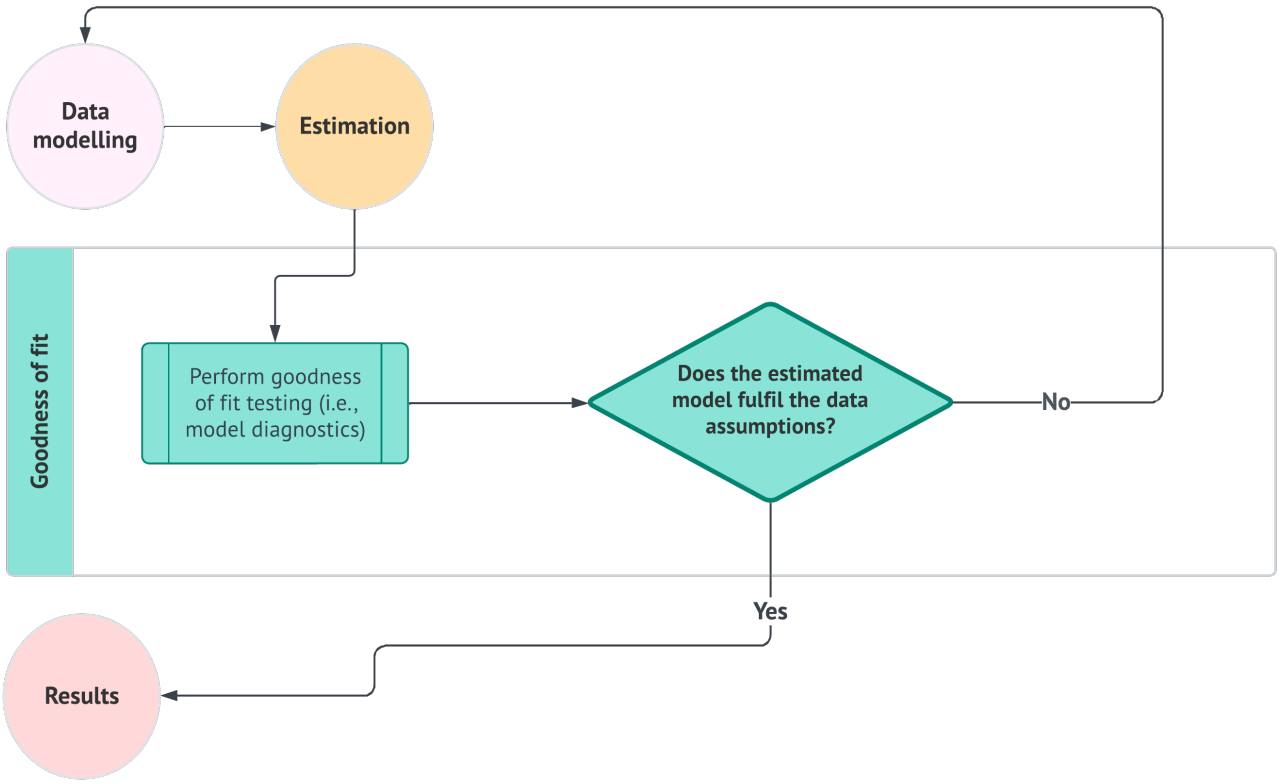

Figure 1.13: *Goodness of fit* stage from the data science workflow in Figure 1.7. This stage is directly preceded by *estimation* and followed by *results*. If necessary, the *goodness of fit* stage could retake the process to *data modelling* and then to *estimation*.

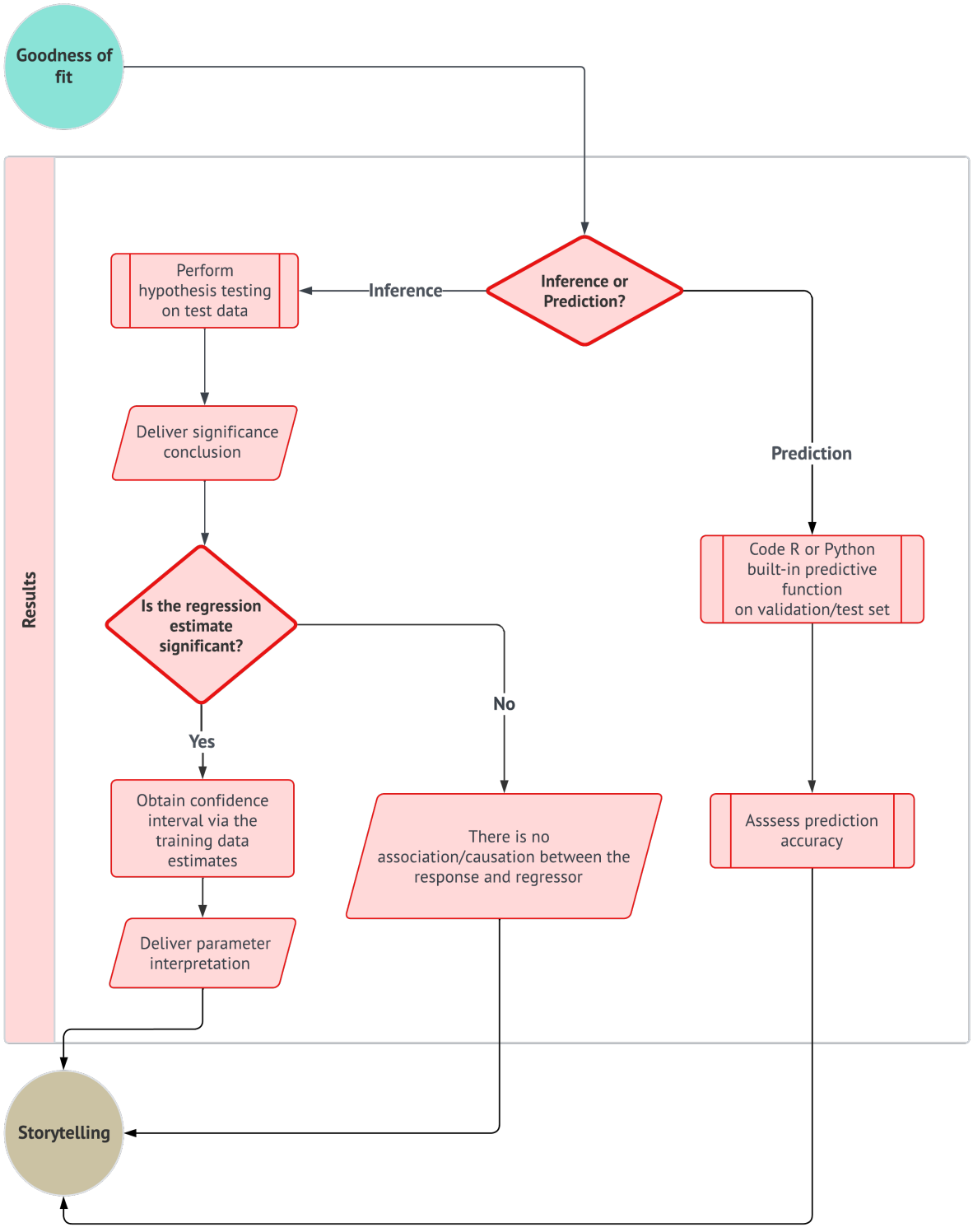

Figure 1.14: *Results* stage from the data science workflow in Figure 1.7. This stage is directly followed by *storytelling* and preceded by *goodness of fit*.

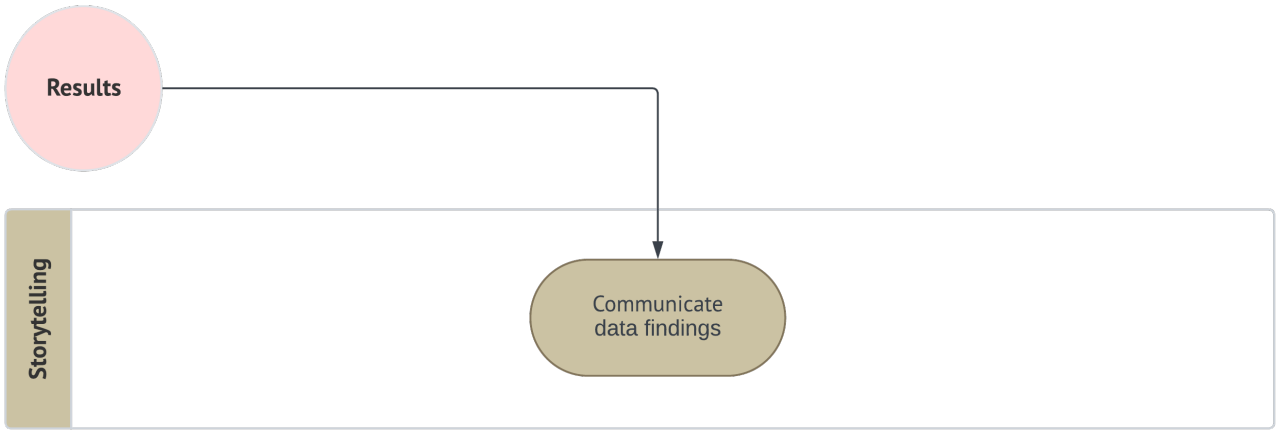

Figure 1.15: *Storytelling* stage from the data science workflow in Figure 1.7. This stage preceded by *results*.

1.3.3 Exploratory Data Analysis

1.3.4 Data Modelling

1.3.5 Estimation

1.3.6 Goodness of Fit

1.3.7 Results

1.3.8 Storytelling

1.4 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 1.16. 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.

That said, we will go beyond OLS regression and explore further regression techniques. In practice, these techniques have been developed in the statistical literature to address practical

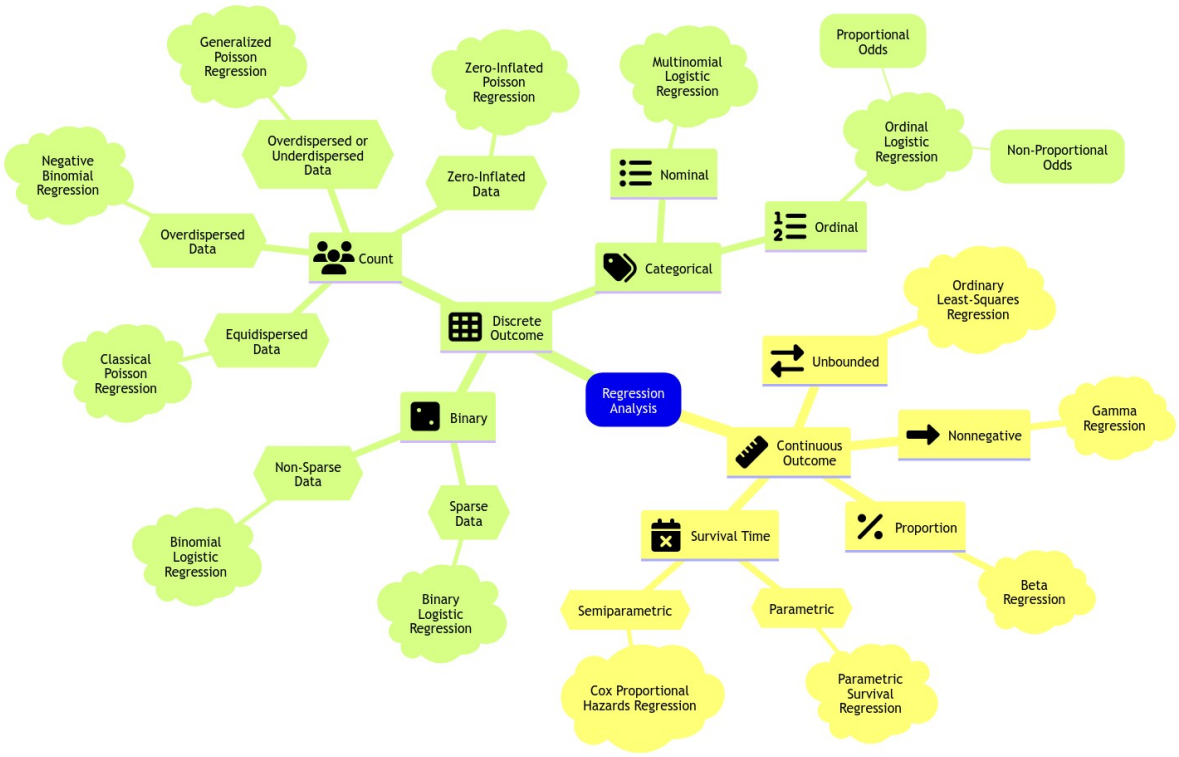

Figure 1.16: Regression analysis mindmap depicting all modelling techniques to be explored in this book. These techniques are split into two big sets: *continuous* and *discrete* outcomes.

cases where the OLS modelling framework and assumptions are not suitable anymore. Thus, throughout this block, we will cover (at least) one new regression model per lecture.

As we can see in the clouds of Figure 1.16, 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 Chapter 2. 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.

Even though this book comprises 13 chapters, each depicting a different regression model, we have split these chapters into two major subsets: those with *continuous* outcomes and those with *discrete* outcomes.

2 Ordinary Least-squares

References

Gelbart, Michael. 2017. "Data Science Terminology." *UBC MDS*. Master of Data Science at the University of British Columbia. https://ubc-mds.github.io/resources_pages/terminology/.

R Core Team. 2024. "R: A Language and Environment for Statistical Computing." Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

The Pandas Development Team. 2024. "Pandas-Dev/Pandas: Pandas." Zenodo. https://doi.org/10.5281/zenodo.3509134.

Van Rossum, Guido, and Fred L. Drake. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

A The ML-Stats Dictionary

Machine learning and statistics comprise a **substantial synergy** that is reflected in data science. Thus, it is imperative to construct solid bridges between both disciplines to ensure everything is clear regarding their tremendous amount of jargon and terminology. This **ML-Stats dictionary** (*ML* stands for *Machine Learning*) aims to be one of these bridges in this textbook, especially within supervised learning and regression analysis contexts.

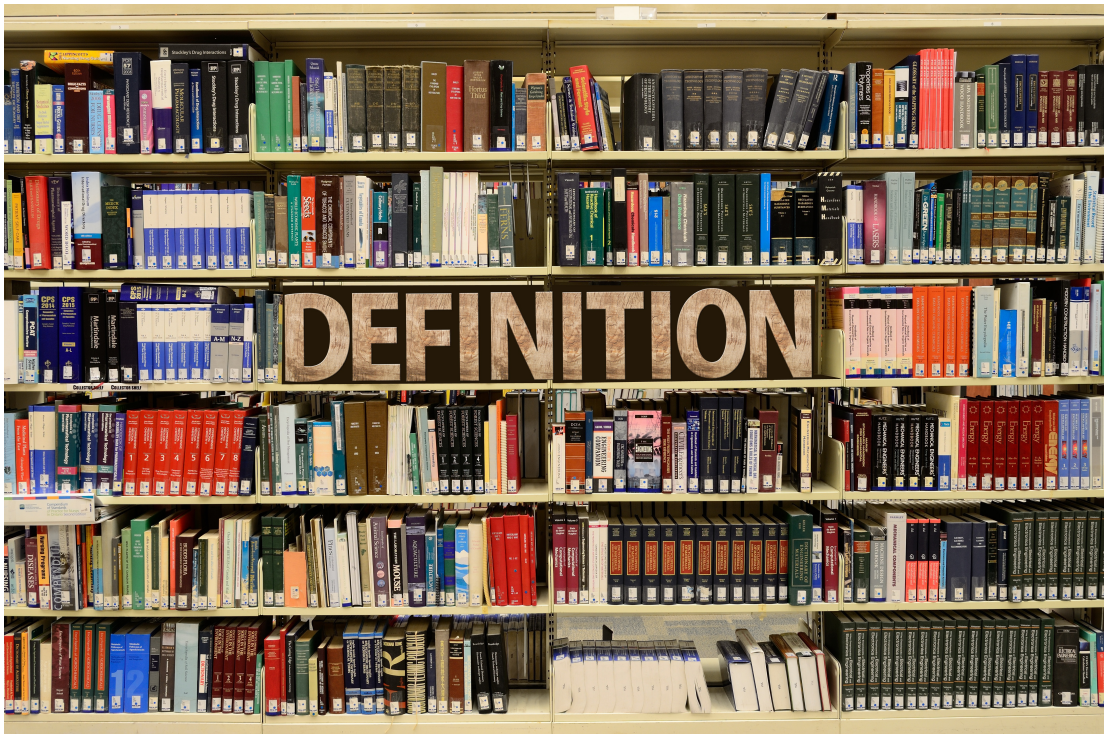

Figure A.1: Image by *Gerd Altmann* via *Pixabay*.

Below, you will find definitions either highlighted in magenta if they correspond to statistical terminology or orange if the terminology is machine learning-related. These definitions come from all **definition** admonitions, such as in Important 1. This colour scheme strives to combine all terminology to switch from one field to another easily. With practice and time, we should be able to jump back and forth when using these concepts.

⚠ Attention!

Noteworthy terms (either statistical or machine learning-related) will include a particular admonition identifying which terms (again, either statistical or machine learning-related) are **equivalent** (**or NOT equivalent if that is the case!**).

D

Dependent variable

In supervised learning, it is the main variable of interest we are trying to **learn** or **predict**, or equivalently, the variable we are trying **explain** in a statistical inference framework.

⚠ Equivalent to:

Response, outcome, output or target.

O

Outcome

In supervised learning, it is the main variable of interest we are trying to **learn** or **predict**, or equivalently, the variable we are trying **explain** in a statistical inference framework.

⚠ Equivalent to:

Dependent variable, response, output or target.

Output

In supervised learning, it is the main variable of interest we are trying to **learn** or **predict**, or equivalently, the variable we are trying **explain** in a statistical inference framework.

⚠ Equivalent to:

Dependent variable, response, outcome or target.

P

Parameter

It is a characteristic (**numerical** or even **non-numerical**, such as a **distinctive category**) that **summarizes** the state of our population or **system** of interest. Examples of a population parameter can be described as follows:

- *The average weight of children between the ages of 5 and 10 years old in states of the American West Coast.*
- *The variability in the height of the mature açaí palm trees from the Brazilian Amazonian jungle.*
- *The proportion of defective items in the production of cellular phones in a set of manufacturing facilities.*
- *The average customer waiting time to get their order in the Vancouver franchises of a well-known ice cream parlour.*
- *The most favourite pizza topping of vegetarian adults between the ages of 30 and 40 years old in Edmonton.*

Note the **standard mathematical notation** for a population parameters are **Greek letters**. Moreover, in practice, these population parameter(s) of interest will be **unknown** to the data scientist or researcher. Instead, they would use formal statistical inference to **estimate** them.

Population

It is a **whole collection of individuals or items** that share **distinctive attributes**. As data scientists or researchers, we are interested in studying these attributes, which we assume are **governed** by parameters. In practice, we must be **as precise as possible** when defining our given population such that we would frame our entire data modelling process since its very early stages. Examples of a population could be the following:

- *Children between the ages of 5 and 10 years old in states of the American west coast.*
- *Customers of musical vinyl records in the Canadian provinces of British Columbia and Alberta.*
- *Avocado trees grown in the Mexican state of Michoacán.*
- *Adult giant pandas in the Southwestern Chinese province of Sichuan.*
- *Mature açaí palm trees from the Brazilian Amazonian jungle.*

Note that the term population could be exchanged for the term **system**, given that certain contexts do not specifically refer to individuals or items. Instead, these contexts could refer to **processes** whose attributes are also governed by parameters. Examples of a **system** could be the following:

- *The production of cellular phones in a set of manufacturing facilities.*
- *The sale process in the Vancouver franchises of a well-known ice cream parlour.*
- *The transit cycle of the twelve lines of Mexico City's subway.*

Probability

Let A be an event of interest in a random phenomenon, in a population or **system** of interest, whose all possible outcomes belong to a given sample space S. Generally, the probability for this event A happening can be mathematically depicted as $P(A)$. Moreover, **suppose we observe the random phenomenon n times** such as we were running some class of experiment, then $P(A)$ is defined as the following ratio:

$$P(A) = \frac{\text{Number of times event } A \text{ is observed}}{n}, \tag{A.1}$$

as the n times we observe the random phenomenon goes to infinity.

Equation A.1 will always put $P(A)$ in the following numerical range:

$$0 \leq P(A) \leq 1.$$

R

Response

In supervised learning, it is the main variable of interest we are trying to **learn** or **predict**, or equivalently, the variable we are trying **explain** in a statistical inference framework.

> ⚠ Equivalent to:
>
> Dependent variable, outcome, outpur or target.

S

Sample space

Let A be an event of interest in a random phenomenon in a population or **system** of interest. The sample space S of event A denotes the set of all the possible **random outcomes** we

might encounter every time we randomly observe A such as we were running some class of experiment.

Note each of these outcomes has a determined probability associated with them. If we add up all these probabilities, the probability of the sample S will be one, i.e.,

$$P(S) = 1. \tag{A.2}$$

T

Target

In supervised learning, it is the main variable of interest we are trying to **learn** or **predict**, or equivalently, the variable we are trying **explain** in a statistical inference framework.

> ⚠ Equivalent to:
>
> Dependent variable, response, outcome or output.

B Greek Alphabet

Statistical notation can be pretty particular and different from **usual mathematical notation**. One of these particularities is the constant use of **Greek letters** to denote unknown population parameters in **modelling setup**, **estimation**, and **statistical inference**. In that spirit, throughout this book, we use diverse Greek letters to denote our regression parameters across each of the outlined models in every chapter.

Figure B.1: Image by *meineresterampe* via *Pixabay*.

During early learning stages of regression modelling, we may feel overwhelmed by these new letters, which could be unfamiliar. Therefore, whenever confusion arises in any of the main chapters in this book regarding the names of these letters, we recommend checking out the Greek alphabet from table Table B.1. Note that frequentist statistical inference mostly uses lowercase letters. **Finally, with practice over time, you would likely end up memorizing most of this alphabet.**

Table B.1: Greek alphabet composed of 24 letters, from *left* to *right* you can find the *name* of letter along with its corresponding *uppercase* and *lowercase* forms.

Name	Uppercase	Lowercase
Alpha	A	α
Beta	B	β
Gamma	Γ	γ
Delta	Δ	δ
Epsilon	E	ϵ
Zeta	Z	ζ
Eta	H	η
Theta	Θ	θ
Iota	I	ι
Kappa	K	κ
Lambda	Λ	λ
Mu	M	μ
Nu	N	ν
Xi	Ξ	ξ
O	O	o
Pi	Π	π
Rho	P	ρ
Sigma	Σ	σ
Tau	T	τ
Upsilon	Υ	υ
Phi	Φ	ϕ
Chi	X	χ
Psi	Ψ	ψ
Omega	Ω	ω