1.1 Introduction to regression analysis

What is regression analysis?

Regression Analysis investigates the functional relationship between statistical variables. Data are usually **multiple observations of a random vector** $(\mathrm{Y}, \mathbf{x})$.

- $\mathbf{x} = (\mathrm{X}_1, \ldots, \mathrm{X}_p)^\top$ is a p-vector of variables termed: *explanatory variables*, regressors, predictors, input variables or *independent variables*.
- Y is called: *response variable*, target variable, output variable, outcome variable or *dependent variable*. It may be continuous ($\in \mathbb{R}$), discrete ($\in \{1, \ldots, K\}$) or ordinal (ordered discrete).

Response variables are usually treated as **random variables**, while predictors are treated as **fixed observations**.

Response and explanatory variables

Response and explanatory variables are measures on one of the following scales:

- **nominal:** when Y is classified into categories, which can be only two (*binary outcome*) or several (*multinomial outcome*)
- **ordinal:** when Y is recorded in classes
- **continuous:** when Y is measured on a continuous scale, at least in theory.

Nominal and ordinal data are discrete variables and can be *qualitative* or *quantitative* (e.g. **counts**). Continuous data are *quantitative*.

$\mathbf{x}$ can also be quantitative or qualitative. In particular, when the explanatory variable is qualitative, it is often called *factor*. A quantitative explanatory variables is called *covariate*.

Response	**Explanatory**	**Method**
Continuous	Binary Nominal, $>$ 2categories Ordinal Continuous	t-test ANOVA ANOVA Multiple regression
Binary	Categorical Continuous	Contingency tables Logistic or probit regression
Nominal, $>$ 2categories	Nominal Categorical & Continuous	Contingency tables Nominal logistic regression
Ordinal	Categorical & Continuous	Ordinal logistic regression
Counts	Categorical Categorical & Continuous	Log-linear models Poisson regression

Regression

We aim to find a "good" functional relationship of the form $\mathrm{Y} = f(\mathbf{x}) + \varepsilon$, where ε is a random error term independent of $\mathbf{x}$ with mean zero.

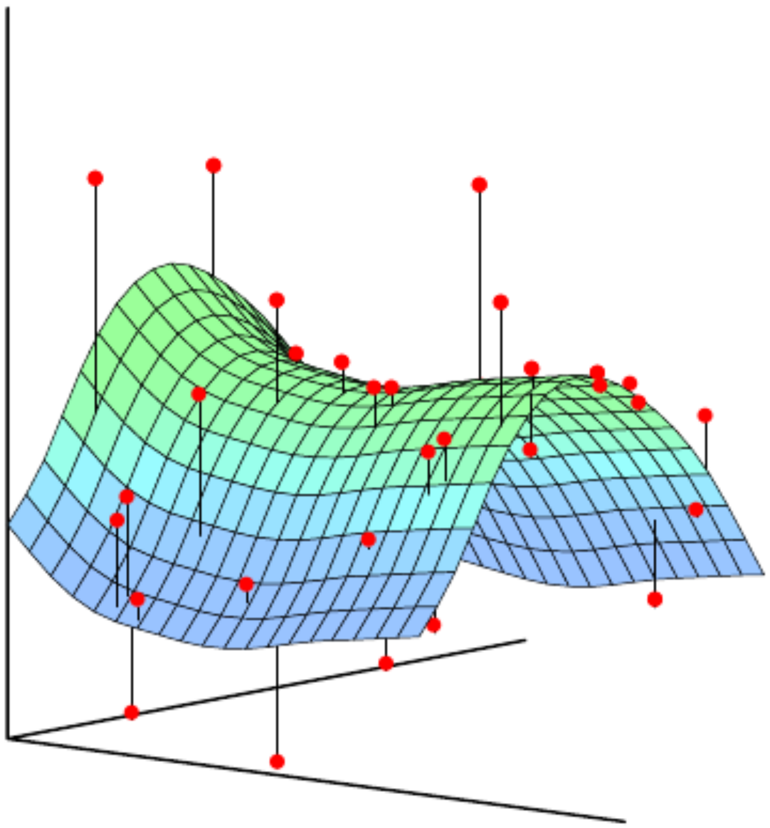

Figure 1.1.1: Regression of Y (vertical, continuous) on $(\mathrm{X}_1, \mathrm{X}_2)$ (horizontal).

"Applied" regression analysis

Applied means: *"If there is no way to calculate it, we won't talk about it."*

On the other hand, we want to understand the underlying computational methods and algorithms. This will be impossible without understanding the theory.

> "There is nothing more practical than a good theory."
>
> — Kurt Lewin

General framework of statistical learning

Statistical learning refers to a vast set of tools for understanding data. It splits into *supervised* and *unsupervised* methods. All the methods presented in this course are within the framework of supervised learning.

Regression fits into the framework of *supervised methods*, which requires a statistical model for predicting or estimating an output based on one or more inputs.

In contrast, *unsupervised methods* cover situations where there are inputs but no supervising output. In these type of analysis we learn about relationships and structure of data. Example of unsupervised analysis is **cluster analysis**.

Knowledge assumed

You need to have a fairly good understanding of **linear algebra**:

- vector spaces,
- linear independence,
- matrix multiplication,
- diagonalisation,
- projections,
- ...

Need to know **multivariate calculus**:

- partial derivatives,
- critical points,
- integrals,
- ...

It's also good to have some previous exposure to **computational software** (`R`):

- data types,
- manipulation of arrays,
- some idea of optimisation,
- ...

And finally, you need to know some **probability theory and statistics**:

- important distributions (normal, Poisson, ...),
- conditional probability,
- conditional expectation,
- covariance matrix,
- asymptotic normality of maximum likelihood estimators,
- ...