Statistical Methods: Likelihood, Bayes and Regression

MATH20802
Lecture Notes

Korbinian Strimmer

University of Manchester

24 February 2021

Contents

Preface

About these notes

This is the course text for MATH20802, an introductory course in **Statistical Methods** for second year mathematics students.

These notes will be updated from time to time. To view the current version in your browser visit the online MATH20802 lecture notes. You may also download the MATH20802 lecture notes as PDF.

About the author

My name is Korbinian Strimmer and I am a Professor in Statistics in the Statistics group of the Department of Mathematics at the University of Manchester. You can find more information about me on my home page.

I have first taught this module in the Spring term 2019 at the University of Manchester.

I hope you enjoy the course. If you have any questions, comments, or corrections then please email me at korbinian.strimmer@manchester.ac.uk

About the module

Topics covered

The MATH20802 module is designed to run over the course of 11 weeks. It has three parts:

1. Likelihood estimation and likelihood ratio tests (W1–W5)
2. Bayesian learning and inference (W6–W8)
3. Linear regression (W9–W11)

This module focuses on conceptual understanding and methods, not on theory, As such, the presentation in this course is non-technical. The aim is to offer insights how diverse statistical approaches are linked and to demonstrate that statistics offers a concise and coherent theory of information rather than being an adhoc collection of "recipes" for data analysis (a common but wrong perception of statistics).

Prerequisites

For this module it is important that you refresh your knowledge in:

- Introduction to statistics
- Probability
- R data analysis and programming

In addition you will need to know matrix algebra and how to compute the gradient and the curvature of a function of several variables.

Check the Appendix for a brief refresher of the essential material.

Additional support material

Accompanying these notes are

- a weekly learning plan for an 11 week study period,
- corresponding worksheets with examples,
- lecture videos (visualiser style).

Organisatorical information is available from the course home page and on Blackboard.

If you are a University of Manchester student and enrolled for this module you can find the exam questions of previous years (without solution) as well as the info about the midterm test on Blackboard.

Furthermore, there is also an MATH20802 online reading list hosted by the University of Manchester library.

Acknowledgements

Many thanks to Beatriz Costa Gomes for her help in creating the 2019 version of the lecture notes and to Kristijonas Raudys for his extensive feedback on the 2020 version.

Part I

Likelihood estimation and inference

Chapter 1

Overview of statistical learning

1.1 How to learn from data?

A fundamental question is how to extract information from data in an optimal way, and to make predictions based on this information.

For this purpose, a number of competing **theories of information** have been developed. **Statistics** is the oldest science of information and is concerned with offering principled ways to learn from data and to extract and process information using probabilistic models. However, there are other theories of information (e.g. Vapnik-Chernov theory of learning, computational learning) that are more algorithmic than analytic and sometimes not even based on probability theory.

Furthermore, there are other disciplines, such computer science and machine learning that are closely linked with and also have substantial overlap with statistics. The field of "data science" today comprises both statistics and a machine learning and brings together mathematics, statistics and computer science. Also the growing field of so-called "artificial intelligence" makes substantial use of statistical and machine learning techniques.

The recent popular science book "The Master Algorithm" by Domingos (2015) provides an accessible informal overview over the various schools of science of information. It discusses the main algorithms used in machine learning and statistics:

- Starting as early as 1763, the **Bayesian school** of learning was started which later turned out to be closely linked with *likelihood inference* established by R.A. Fisher in 1922 and generalised in 1951 to **entropy learning** by Kullback and Leibler.
- It was also in the 1950s that the concept of artificial **neural network** arises, essentially a nonlinear input-output map that works in a non-probabilistic way. This field saw another leap in the 1980 and further progress from

2010 onwards with the development of *deep dearning*. It is now one of the most popular (and most effective) methods for analysis of imaging data. Even your mobile phone most likely has a dedicated computer chip with special neural network hardware, for example.

- Further advanced theories of information where developed in the 1960 under the term of **computational learning**, most notably the Vapnik-Chernov theory, with the most prominent example of the "support vector machine" (another non-probabilistic model).
- With the advent of large-scale genomic and other high-dimensional data there has been a surge of new and exciting developments in the field of high-dimensional (large dimension) and also big data (large dimension and large sample size), both in statistics and in machine learning.

The connections between various fields of information is still not perfectly understood, but it is clear that an overarching theory will need to be based on probabilistic learning.

1.2 Probability theory versus statistical learning

When you study statistics (or any other information theory) you need to be aware that there is a fundamental difference between probability theory and statistics, and that relates to the **distinction between "randomness" and "uncertainty"**.

Probability theory studies **randomness**, by developing mathematical models for randomness (such as probability distributions), and studying corresponding mathematical properties (including asymptotics etc). Probability theoy may in fact be viewed as a branch of measure theory, and thus it belongs to the domain pure mathematics.

Probability theory provides probabilistic generative models for data, for simulation of data or for use in learning from data, i.e. inference about the model from observations. Methods and theory how to best learn from data is the domain of applied mathematics, specifically statistics and the related areas of machine learning and data science.

Note that statistics, in contrast to probability, is in fact not at all concerned with randomness. Instead, the focus is about measuring and elucidating the **uncertainty** of events, predictions, outcomes, parameters and this uncertainty measures the **state of knowledge**. Note that if new data or information becomes available, the state of knowledge and thus the uncertainty changes! Thus, **uncertainty is an epistemological property**.

The uncertainty most often is due to our ignorance of the true underlying processes (on purpose or not), but not because the underlying process is actually random. The success of statistics is based on the fact that we can mathematically model the uncertainty without knowing any specifics of the underlying processes, and we still have procedures for optimal inference under uncertainty.

In short, statistics is about describing the state of knowledge of the world, which may be uncertain and incomplete, and to make decisions and prediction in the

face of uncertainty, and this uncertaintly sometimes derives from randomness but most often from our ignorance (and sometimes this ignorance even helps to create a simple yet effective model)!

1.3 Cartoon of statistical learning

We observe data $x_1, \ldots, x_n$ assumed to be generated by the underlying true model M_{true}.

To explain the data, and make prediction, we make hypotheses in the form of candidate models $M_1, M_2, \ldots$. The true model M_{true} itself is unknown and cannot be observed. However, what we can observe is a finite amount of data from the model by measuring properties of objects interest (our observations from experiments). Sometimes we can also perturb the model and see what the effect is (interventional study).

The various candidate models $M_1, M_2, \ldots$ in the **model world** will never be perfect or correct as the true model M_{true} will only be among the candidate models in an idealised situation. However, even an imperfect candidate model will often provide a useful mathematical approximation and capture some important characteristics of the true model and thus will help to interpret observed data..

$$\begin{array}{ccccc} & & & & \textbf{Model world} \\ & \textbf{Hypothesis} & & & M_1 : f(x|\theta_1) \\ & \text{How the world works} & \longrightarrow & & M_2 : f(x|\theta_2) \\ & & & & \vdots \\ & & & & \\ & & \textbf{Real world,} & & \\ \longrightarrow & & \textbf{unknown true model} & \longrightarrow & \textbf{Data } x_1, \ldots, x_n \\ & & M_{\text{true}} : f(x|\theta_{\text{true}}) & & \end{array}$$

The aim of statistical learning is to identify the model(s) that explain the current data and also predict future data (i.e. predict outcome of experiments that have not been conducted yet).

Thus a good model provides a good fit to the current data (i.e. it explains current observations well) and also to future data (i.e. it generalises well).

A large proportion of statistical theory is devoted to finding these "good" models that avoid both *overfitting* (models being too complex and don't generalise well) or *underfitting* (models being too simplistic and hence also don't predict well).

Typically the aim is to find a model whose the **model complexity** matches the complexity of the unknown true model and also the complexity of the data observed from the unknown true model.

1.4 Likelihood

A core problem in statistics is how to find probabilistic models for explaining existing data and predicting new data. For this we need a measure of how

good a hypothesis/candidate model M_k is as approximation for the (typically unknown) true data generating model M_{true}.

As you already know from the year 1 module MATH10282 "Introduction to Statistics", one such measure is provided by the likelihood function which helps to choose among the various candidate models and estimate corresponding parameters by finding the model M that maximises the (log)-likelihood.

Given a probability distribution $F_{\boldsymbol{\theta}}$ with density or mass function $f(x|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is a parameter vector, and $x_1, \ldots, x_n$ is the observed data (independent and identically distributed), the **likelihood function** is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})$$

Typically, instead of the likelihood one uses the log-likelihood function:

$$\log L(\boldsymbol{\theta}) = l_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(x_i|\boldsymbol{\theta})$$

Reasons for using log-likelihood (rather than likelihood) include that

- the log density is in fact the more "natural" and relevant quantity (this will become clear in the upcoming chapters) and that
- addition is numerically more stable than multiplication on a computer.

For discrete random variables where $f(x|\boldsymbol{\theta})$ is a probability mass function the likelihood can be interpreted as the probability to observe the data given the model with specified parameters $\boldsymbol{\theta}$. However, for continuous random variables this interpretation breaks down.

In the next chapter we will see that the justification for using likelihood rather stems from its close link to the Kullback-Leibler information. This also helps to understand why using likelihood for estimation is only optimal in the limit of large sample size.

In the first part of the MATH28082 "Statistical Methods" module we will study likelihood estimation and inference in much detail. We will provide links to related methods of inference and discuss its information-theoretic foundations. We will also discuss the optimality properties as well as the limitation of likelihood inference. Extensions of likelihood analysis, in particular Bayesian learning, which will be discussed in the second part module. In the third part of the module we will apply statistical learning to linear regression.

Chapter 2

From information theory to likelihood

2.1 Entropy

2.1.1 Overview

In this chapter we discuss various information criteria and their connection to maximum likelihood.

The modern definition of (relative) entropy, or "disorder", was first discovered by physicist Boltzmann in 1875 in the context of thermodynamics. In the 1940-1950's the notion of entropy turned out to be central in information theory, a field pioneered by mathematicians such as Hartley, Good, Jaynes, Kullback, Leibler, Shannon, and Turing, and later further explored by Amari, Ciszar, Dawid, Efron and many others.

Entropy $\nearrow$ Shannon Entropy (Shannon 1948)

Entropy $\searrow$ Relative Entropy (Kullback-Leibler 1951)

Fisher information $\rightarrow$ Likelihood theory (Fisher 1922)

Mutual Information $\rightarrow$ Information theory (Shannon 1948, Lindley 1953)

2.1.2 Surprise

We assume a *discrete* random variable x with state space $\Omega = \{x_1, x_2, \ldots, x_K\}$ with finite K and an associated distribution F. The corresponding probability mass function (pmf) is given by $f(x_i) = \Pr(x = x_i) = p_i$, with each probability

$p_i \in [0,1]$ and $\sum_{i=1}^{K} p_i = 1$. F is known as **categorical distribution** with K classes. For $K = 2$ it reduces to the **Bernoulli distribution**.

The **surprise** to observe x_i is then defined as $-\log(p_i)$. Thus, the surprise to observe a certain event (with $p_i = 1$) is zero, and conversely the surprise to observe an event that is certain not to happen (with $p_i = 0$) is infinite.

Note that in this module we always use the *natural logarithm* by default, and will explicitly write $\log_2$ and $\log_{10}$ for logarithms with respect to base 2 and 10, respectively.

Surprise and entropy computed with the natural logarithm (log) is given in "nats" (=natural information units. If it is computed using $\log_2$ then the units are "bits", and in terms of $\log_{10}$ the units is "ban" or "Hartley", cf. https://en.wikipedia.org/wiki/Hartley_(unit) .

2.1.3 Shannon entropy

The **Shannon** entropy of the distribution F is defined as the **expected surprise**, i.e. the negative expected log-probability

$$H(F) = -\mathrm{E}_F\left(\log f(x)\right) = -\sum_{i=1}^{K} f(x_i)\log f(x_i) = -\sum_{i=1}^{K} p_i \log(p_i)$$

Often the entropy $H(F)$ is written as $H(x)$, but the first notation is less ambiguous especially if more than one distribution is considered.

As $p \in [0,1]$ the term $-p\log(p)$ equals zero only for $p = 0$ or $p = 1$ and is otherwise positive. As a consequence, Shannon entropy is bounded below by zero. If K is finite Shannon entropy is also bounded above by $\log K$ and therefore

$$\log K \geq H(F) \geq 0$$

for any discrete distribution F with K categories.

Example 2.1. Discrete uniform distribution U_K: let $p_1 = p_2 = \ldots = p_K = \frac{1}{K}$. Then

$$H(U_k) = -\sum_{i=1}^{K} \frac{1}{K}\log\left(\frac{1}{K}\right) = \log K$$

Note this is the largest value the Shannon entropy can assume with K classes.

Example 2.2. Concentrated probability mass: let $p_1 = 1$ and $p_2 = p_3 = \ldots = p_K = 0$. Using $0 \times \log(0) = 0$ we obtain for the Shannon probability

$$H(F) = 1 \times \log(1) + 0 \times \log(0) + \cdots = 0$$

Note that 0 is the smallest value that Shannon entropy can assume, and corresponds to maximum concentration.

Thus, **large entropy** implies that the **distribution is spread out** whereas **small entropy** means the **distribution is concentrated**.

Correspondingly, maximum entropy distributions can be considered minimally informative about a random variable.

Multivariate case:

If $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$ is a multivariate random variable of dimension d with distribution F then $H(F)$ is called the *joint entropy* of the random variables $x_1, x_2, \ldots, x_d$. It is also often written as $H(x_1, x_2, \ldots, x_d)$.

2.1.4 Differential entropy

Shannon entropy is only defined for discrete random variables.

Differential Entropy results from applying the definition of Shannon entropy to a *continuous* random variable x with density $f(x)$:

$$H(F) = -\mathrm{E}_F(\log f(x)) = -\int f(x) \log f(x)\, dx$$

Despite having essentially the same formula the different name is justified because differential entropy exhibits different properties compared to Shannon entropy, because the logarithm is taken of a density which in contrast to a probability can assumes values larger than one. As a consequence, differential entropy is *not* bounded below by zero and can be negative.

Example 2.3. Consider the uniform distribution $U(0, a)$ with $a > 0$, support from 0 to a and density $f(x) = 1/a$. As $-\int_0^a f(x) \log f(x) dx = -\int_0^a \frac{1}{a} \log(\frac{1}{a}) dx = \log a$ the differential entropy is

$$H(U(0, a)) = \log a\,.$$

Note that for $a < 1$ the differential entropy is negative.

Example 2.4. The density of the univariate normal $N(\mu, \sigma^2)$ distribution is $f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ with $\sigma^2 > 0$. The corresponding differential entropy is is

$$H(F) = \frac{1}{2}(\log(2\pi\sigma^2) + 1)\,.$$

Note that it only depends on the variance and not on the mean, and that for $\sigma^2 < 1/(2\pi e) \approx 0.0585$ the differential entropy is negative.

2.1.5 Maximum entropy principle to characterise distributions

Both maximum Shannon entropy and differential entropy are useful to characterise distributions:

1) The **discrete uniform distribution** is the **maximum entropy distribution** among all categorical distributions.

2) the maximum entropy distribution of a continuous random variable with support $[-\infty, \infty]$ with a specific mean and variance is the normal distribution

3) the maximum entropy distribution among all continuous distributions supported in $[0, \infty]$ with a specified mean is the exponential distribution.

The higher the entropy the more spread out (and more uninformative) is a distribution.

Using maximum entropy to characterise maximally uniformative distributions was advocated by E.T. Jaynes (who also proposed to use maximum entropy in the context of finding Bayesian priors).

A list of maximum entropy distribution is given here: https://en.wikipedia.org/wiki/Maximum_entropy_probability_distribution .

2.1.6 Cross-entropy

If in the definition of Shannon entropy (and differential entropy) the expectation over the log-density (say $g(x)$ of distribution G) is with regard to a different distribution F over the same state space we arrive at the **cross-entropy**

$$H(F, G) = -\mathrm{E}_F \left(\log g(x)\right)$$

Therefore, cross-entropy is a measure linking two distributions F and G.

Note that

- cross-entropy is not symmetric with regard to F and G, because the expectation is taken with reference to F.
- By construction $H(F, F) = H(F)$.
- if G is uniform then $H(F, G) = C$ (constant, not depending on F).

A crucial property of the cross-entropy $H(F, G)$ is that it is bounded below by the entropy of F, therefore

$$H(F, G) \geq H(F)$$

with equality for $F = G$.

Equivalently we can write

$$H(F, G) - H(F) \geq 0$$

In fact, this recalibrated cross-entropy turns out to be more fundamental than both cross-entropy and Shannon resp. differential entropy. It will be studied in detail in the next section.

Example 2.5. Cross-entropy between two normals:

Assume $F_{\text{ref}} = N(\mu_{\text{ref}}, \sigma^2_{\text{ref}})$ and $F = N(\mu, \sigma^2)$. Then the cross-entropy is

$$H(F_{\text{ref}}, F) = \frac{1}{2}\left(\frac{(\mu - \mu_{\text{ref}})^2}{\sigma^2} + \frac{\sigma^2_{\text{ref}}}{\sigma^2} + \log(2\pi\sigma^2)\right)$$

Example 2.6. If $\mu_{\text{ref}} = \mu$ and $\sigma^2_{\text{ref}} = \sigma^2$ then the above cross-entropy $H(F, G)$ degenerates to the differential entropy $H(F_{\text{ref}}) = \frac{1}{2}\left(\log(2\pi\sigma^2_{\text{ref}}) + 1\right)$.

2.2 Kullback-Leibler divergence

2.2.1 Definition

Also known as **relative entropy** and **discrimination information**.

The **relative entropy** measures the **divergence** of a distribution G from the distribution F and is defined as

$$\begin{aligned} D_{\mathrm{KL}}(F,G) &= \mathrm{E}_F \log\left(\frac{f(x)}{g(x)}\right) \\ &= \underbrace{-\mathrm{E}_F(\log g(x))}_{\text{cross-entropy}} - (\underbrace{-\mathrm{E}_F(\log f(x))}_{\text{(differential) entropy}}) \\ &= H(F,G) - H(F) \end{aligned}$$

- $D_{\mathrm{KL}}(F,G)$ measures the amount of information lost if G is used to approximate F.
- If F and G are identical (and no information is lost) then $D_{\mathrm{KL}}(F,G) = 0$.
- If G is uniform then $D_{\mathrm{KL}}(F,G) = -H(F) + C$, i.e. the negative Shannon/differential entropy of F measures the loss when F is approximated by a uniform.

(Note: here "divergence" measures the dissimilarity between probability distributions. This type of divergence is not related and should not be confused with divergence (div) as used in vector analysis.)

The term divergence (rather than distance) implies also that the distributions F and G are not interchangeable in $D_{\mathrm{KL}}(F,G)$.

In applications in statistics the typical roles of F and G are:

- F as the (unknown) underlying true model for the data generating process
- G as the approximating model (e.g. some parametric family)

In Bayesian statistics we use

- F as posterior distribution
- G as prior distribution

There exist various notations for KL divergence in the literature. Here we use $D_{KL}(F,G)$ but often you can find $KL(F||G)$ or $\boldsymbol{I}^{KL}(F;G)$ in other references.

2.2.2 Properties of KL divergence

1. $D_{\mathrm{KL}}(F,G) \neq D_{\mathrm{KL}}(G,F)$, i.e., KL divergence is not symmetric, F and G cannot be interchanged.
2. $D_{\mathrm{KL}}(F,G) = 0$ if and only if $F = G$, i.e., the KL divergence is zero if and only if F and G are identical.
3. $D_{\mathrm{KL}}(F,G) \geq 0$, proof via the **Jensen Inequality**.
4. $D_{\mathrm{KL}}(F,G)$ remains invariant under coordinate transformations, i.e. it is an invariant geometric quantity.

Note that in the KL divergence the expectation is taken over a ratio of densities (or ratio of probabilities for discrete random variables). This is what creates the transformation invariance.

For more details and proofs of properties 3 and 4 see Worksheet 1.

2.2.3 Examples

Example 2.7. KL divergence between two univariate normals:

Assume $F_{\text{ref}} = N(\mu_{\text{ref}}, \sigma^2_{\text{ref}})$ and $F = N(\mu, \sigma^2)$. Then

$$\begin{aligned} D_{\text{KL}}(F_{\text{ref}}, F) &= H(F_{\text{ref}}, F) - H(F_{\text{ref}}) \\ &= \frac{1}{2}\left(\frac{(\mu - \mu_{\text{ref}})^2}{\sigma^2} + \frac{\sigma^2_{\text{ref}}}{\sigma^2} - \log\left(\frac{\sigma^2_{\text{ref}}}{\sigma^2}\right) - 1\right) \end{aligned}$$

As special case if variances are equal $\sigma^2 = \sigma^2_{\text{ref}}$ we get

$$D_{\text{KL}}(F_{\text{ref}}, F) = \frac{1}{2}\left(\frac{(\mu - \mu_{\text{ref}})^2}{\sigma^2_{\text{ref}}}\right)$$

so the KL divergence reduces to the squared standardised difference of means.

Example 2.8. KL divergence between two categorical distributions P and Q:

With probabilities $p_1, \ldots, p_k$ and $q_1, \ldots, q_k$ we get:

$$D_{\text{KL}}(P, Q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right) \approx \begin{cases} \frac{1}{2}\sum_i \frac{(p_i - q_i)^2}{q_i} = \frac{1}{2} D_{\text{Pearson}}(P, Q) & (=p_i \text{ around } q_i) \\ \frac{1}{2}\sum_i \frac{(p_i - q_i)^2}{p_i} = \frac{1}{2} D_{\text{Pearson}}(Q, P) & (=q_i \text{ around } p_i) \end{cases}$$

$D_{\text{Pearson}}(P; Q) = \sum_i \frac{(p_i - q_i)^2}{q_i}$ is called the Pearson χ^2 divergence, a member of the family of f-divergences.

If O and E are given as observed and expected distribution, with corresponding observed frequencies o_i and expected frequencies e_i and n the number of total counts, then $n D_{\text{Pearson}}(O; E) = n \sum_{i=1}^{d} \frac{(o_i - e_i)^2}{e_i} = X^2_{\text{Pearson}}$ yields the Pearson χ^2 test statistic.

Thus, the chi-squared statistic is in effect an second order approximation to the KL divergence between two discrete distributions!

See Worksheet 1 for further details and derivations.

2.3 Local quadratic approximation and expected Fisher information

KL information measures the divergence of two distributions. We may thus use relative entropy to measure the divergence between two distributions in the same family, separated in parameter space only by some small $\boldsymbol{\varepsilon}$:

$$D_{\text{KL}}(F_{\boldsymbol{\theta}}, F_{\boldsymbol{\theta}+\boldsymbol{\varepsilon}}) = ?$$

To evaluate this we first consider the quadratic approximation of the log density $\log f(x|\boldsymbol{\theta}) = h(\boldsymbol{\theta})$ as a function of the parameter vector $\boldsymbol{\theta}$. A Taylor series expansion $h(\boldsymbol{\theta}+\boldsymbol{\varepsilon}) = h(\boldsymbol{\theta}) + \nabla h(\boldsymbol{\theta})\boldsymbol{\varepsilon} + \frac{1}{2}\boldsymbol{\varepsilon}^T\nabla^T\nabla h(\boldsymbol{\theta})\boldsymbol{\varepsilon} + \ldots$ yields in second order

$$\log f(x|\boldsymbol{\theta}+\boldsymbol{\varepsilon}) \approx \log f(x|\boldsymbol{\theta}) + \nabla \log f(x|\boldsymbol{\theta})\,\boldsymbol{\varepsilon} + \frac{1}{2}\boldsymbol{\varepsilon}^T\nabla^T\nabla \log f(x|\boldsymbol{\theta})\boldsymbol{\varepsilon}\,.$$

Note that the gradient and the Hessian matrix is computed with regard to $\boldsymbol{\theta}$ and x is assumed fixed.

With this the KL divergence between $F_{\boldsymbol{\theta}}$ and $F_{\boldsymbol{\theta}+\boldsymbol{\varepsilon}}$ becomes

$$\begin{aligned}
D_{\mathrm{KL}}(F_{\boldsymbol{\theta}}, F_{\boldsymbol{\theta}+\boldsymbol{\varepsilon}}) &= \mathrm{E}_{F_{\boldsymbol{\theta}}} \log f(x|\boldsymbol{\theta}) - \mathrm{E}_{F_{\boldsymbol{\theta}}} \log f(x|\boldsymbol{\theta}+\boldsymbol{\varepsilon}) \\
&\approx -\mathrm{E}_{F_{\boldsymbol{\theta}}} \left(\nabla \log f(x|\boldsymbol{\theta})\,\boldsymbol{\varepsilon}\right) - \mathrm{E}_{F_{\boldsymbol{\theta}}} \left(\frac{1}{2}\boldsymbol{\varepsilon}^T\nabla^T\nabla \log f(x|\boldsymbol{\theta})\boldsymbol{\varepsilon}\right) \\
&= -\underbrace{\mathrm{E}_{F_{\boldsymbol{\theta}}} \left(\nabla \log f(x|\boldsymbol{\theta})\right)}_{\text{vanishes, see below}} \boldsymbol{\varepsilon} + \frac{1}{2}\boldsymbol{\varepsilon}^T \mathrm{E}_{F_{\boldsymbol{\theta}}} \left(-\nabla^T\nabla \log f(x|\boldsymbol{\theta})\right) \boldsymbol{\varepsilon} \\
&= \frac{1}{2}\boldsymbol{\varepsilon}^T \underbrace{\boldsymbol{I}^{\mathrm{Fisher}}(\boldsymbol{\theta})}_{\text{expected Fisher information}} \boldsymbol{\varepsilon}
\end{aligned}$$

The term $\mathrm{E}_{F_{\boldsymbol{\theta}}} \left(\nabla \log f(x|\boldsymbol{\theta})\right) = 0$ because $\nabla \log f(x|\boldsymbol{\theta}) = f(x|\boldsymbol{\theta})^{-1}\nabla f(x|\boldsymbol{\theta})$ and thus $\mathrm{E}_{F_{\boldsymbol{\theta}}} \left(\nabla \log f(x|\boldsymbol{\theta})\right) = \int \nabla f(x|\boldsymbol{\theta}) d\boldsymbol{\theta} = \nabla \int f(x|\boldsymbol{\theta}) d\boldsymbol{\theta} = \nabla 1 = 0$ assuming exchange of integration and differentiation is possible.

Therefore, the expected Fisher information matrix

$$\boldsymbol{I}^{\mathrm{Fisher}} = -\mathrm{E}_{F_{\boldsymbol{\theta}}} \left(\nabla^T\nabla \log f(x|\boldsymbol{\theta})\right)$$

arises from a local quadratic approximation of KL divergence.

Note that as there is no data involved as the expected Fisher information is purely a property of the model, or more precisely of the space of the models indexed by $\boldsymbol{\theta}$. In fact, in *information geometry* the expected Fisher information plays an important role as the metric tensor of this space (https://en.wikipedia.org/wiki/Fisher_information_metric).

Example 2.9. Expected Fisher information for the normal distribution $N(\mu, \sigma^2)$.

The log-density is

$$\log f(x|\mu, \sigma^2) = -\frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(x-\mu)^2 - \frac{1}{2}\log(2\pi)$$

The gradient with respect to μ and σ^2 (!) is

$$\nabla \log f(x|\mu, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2}(x-\mu) \\ -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(x-\mu)^2 \end{pmatrix}$$

Hint for calculating the gradient: replace σ^2 by v and then take the partial derivative with regard to v, then substitute back.

The Hessian matrix is

$$\nabla^T \nabla \log f(x|\mu,\sigma^2) = \begin{pmatrix} -\frac{1}{\sigma^2} & -\frac{1}{\sigma^4}(x-\mu) \\ -\frac{1}{\sigma^4}(x-\mu) & \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}(x-\mu)^2 \end{pmatrix}$$

As $\mathrm{E}(x) = \mu$ we have $\mathrm{E}(x-\mu) = 0$. Furthermore, with $\mathrm{E}((x-\mu)^2) = \sigma^2$ we see that $\mathrm{E}\left(\frac{1}{\sigma^6}(x-\mu)^2\right) = \frac{1}{\sigma^4}$. Therefore the expected Fisher information matrix as the negative expected Hessian matrix is

$$\boldsymbol{I}^{\text{Fisher}}\left(\mu,\sigma^2\right) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$$

2.4 Entropy learning and maximum likelihood

2.4.1 The relative entropy between true model and approximating model

Assume we have observations $x_1, \ldots, x_n$. The data is sampled from F, the true but unknown data generating distribution. We also specify models $G_{\boldsymbol{\theta}}$ indexed by $\boldsymbol{\theta}$ to approximate F.

The relative entropy $D_{\mathrm{KL}}(F, G_{\boldsymbol{\theta}})$ then measures the divergence of the approximation $G_{\boldsymbol{\theta}}$ from the unknow true model F. It can be written as:

$$D_{\mathrm{KL}}(F, G_{\boldsymbol{\theta}}) = \underbrace{-\mathrm{E}_F \log g_{\boldsymbol{\theta}}(x)}_{\text{cross-entropy}} - (\underbrace{-\mathrm{E}_F \log f(x)}_{\text{entropy of } F\text{, does not depend on } \boldsymbol{\theta}})$$

However, since we do not know F we cannot actually compute this divergence. Nonetheless, we may use the empirical distribution $\hat{F}_n$ — a function of the observed data — as approximation for F, and in this way arrive at an approximation for $D_{\mathrm{KL}}(F, G_{\boldsymbol{\theta}})$ that becomes more and more accurate with growing sample size.

Recall "Law of Large Numbers" :

- By the strong law of large numbers the empirical distribution $\hat{F}_n$ converges to the true underlying distribution F as $n \to \infty$ almost surely:

$$\hat{F}_n \overset{a.s.}{\to} F$$

- For $n \to \infty$ the average $\mathrm{E}_{\hat{F}_n}(h(X)) = \frac{1}{n}\sum_{i=1}^{n} h(x_i)$ converges to the expectation $\mathrm{E}_F(h(X))$.

Hence, for large sample size n we can approximate cross-entropy and as a result the KL divergence. The cross-entropy $H(F, G_{\boldsymbol{\theta}})$ is approximated by the

empirical cross-entropy where the expectation is taken with regard to $\hat{F}_n$ rather than F:

$$\begin{aligned} H(F, G_{\boldsymbol{\theta}}) &\approx H(\hat{F}_n, G_{\boldsymbol{\theta}}) \\ &= -\mathrm{E}_{\hat{F}_n}(\log(x)) \\ &= -\frac{1}{n}\sum_{i=1}^{n} \log g(x_i|\boldsymbol{\theta}) \\ &- \frac{1}{n} l_n(\boldsymbol{\theta}) \end{aligned}$$

This turns out to be equal to the negative log-likelihood standardised by the sample size n! Or in other words, the **likelihood** is the **negative empirical cross-entropy (times sample size n)**.

The KL divergence $D_{\mathrm{KL}}(F, G_{\boldsymbol{\theta}})$ can therefore be approximated by

$$D_{\mathrm{KL}}(F, G_{\boldsymbol{\theta}}) \approx -\frac{1}{n} l_n(\boldsymbol{\theta}) + C$$

where C is a constant (the negative entropy of the true distribution F) that does not depend on the parameters $\boldsymbol{\theta}$ (and hence does not matter when optimising the divergence).

2.4.2 Minimum KL divergence and maximum likelihood

If we were to know F we would simply minimise $D_{\mathrm{KL}}(F, G_{\boldsymbol{\theta}})$ to find the particular model $G_{\boldsymbol{\theta}}$ that is closest to the true model. Equivalently, we would minimise the cross-entropy $H(F, G_{\boldsymbol{\theta}})$. However, since we actually don't know F this is not possible.

However, for large sample size n when the empirical distribution $\hat{F}_n$ is a good approximation for F, we can use the above approximation. Thus, instead of minimising the KL divergence $D_{\mathrm{KL}}(F, G_{\boldsymbol{\theta}})$ we simply minimise $H(\hat{F}_n, G_{\boldsymbol{\theta}})$ which is the same as maximising the likelihood $l_n(\boldsymbol{\theta})$.

Conversely, it turns out that maximising the likelihood with regard to the $\boldsymbol{\theta}$ is equivalent (at least asymptotically for large n!) to minimising the KL divergence of the approximating model and the unknown true model!

$$\begin{aligned} \hat{\boldsymbol{\theta}}^{ML} &= \arg\max_{\boldsymbol{\theta}} \; l_n(\boldsymbol{\theta}) \\ &= \arg\min_{\boldsymbol{\theta}} \; H(\hat{F}_n, G_{\boldsymbol{\theta}}) \\ &\approx \arg\min_{\boldsymbol{\theta}} \; D_{\mathrm{KL}}(F, G_{\boldsymbol{\theta}}) \end{aligned}$$

Therefore, the reasoning behind the method of **maximum likelihood** is that it minimises a **large sample approximation of the KL divergence** of the candidate model $G_{\boldsymbol{\theta}}$ from the unkown true model F.

As a consequence of the close link of maximum likelhood and relative entropy maximum likelihood inherits for large n (and only then!) all the optimality

properties from KL divergence. These will be discussed in more detail later in the course.

2.4.3 Further connections

Since minimising KL divergence contains ML estimation as special case you may wonder whether there is a broader justification of relative entropy in the context of statistical data analysis? Indeed, KL divergence has strong geometrical interpretation that forms the basis of *information geometry* and it is also linked stronlgy to to probabilistic forecasting.

In the framework of so-called *scoring rules* (cf. https://en.wikipedia.org/wiki/Scoring_rule) the only local proper scoring rule is the negative log-probability ("surprise"). The expected "surprise" is the cross-entropy and relative entropy is the corresponding natural divergence connected with the log scoring rule.

Furthermore, another intriguing property of KL divergence is that the relative entropy $D_{\mathrm{KL}}(F, G)$ is the *only divergence measure* that is both a Bregman and f-divergence. Note that f-divergences and Bregman-divergences (in turn related to proper scoring rules) are two large classes of measures of similarity and divergence between two probability distributions (see https://en.wikipedia.org/wiki/Bregman_divergence and https://en.wikipedia.org/wiki/F-divergence) Finally, not only the likelihood estimation but also the Bayesian update rule (as discussed later in this module) is another special case of entropy learning.

Chapter 3

Maximum likelihood estimation

3.1 Principle of maximum likelihood estimation

3.1.1 Outline

This chapter discusses maximum likelihood (ML) estimation, in particular ML point estimation, and provides a number of worked examples of ML estimators.

Starting point:

- observed data $x_1, \ldots, x_n$
- model $F_{\boldsymbol{\theta}}$ with density or mass function $f(x|\boldsymbol{\theta})$ with parameters $\boldsymbol{\theta}$

Likelihood function:

- $L(\boldsymbol{\theta}|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})$

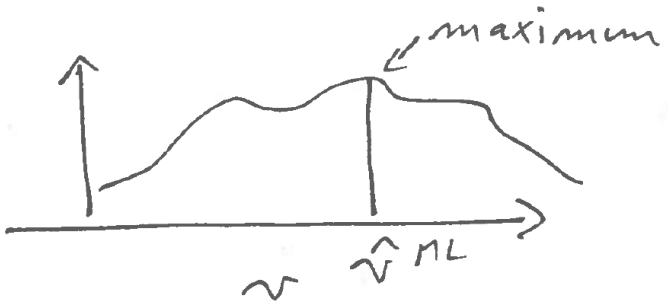
The likelihood is a large sample approximation to the cross-entropy between the unknown true data generating model and the approximating model $F_{\boldsymbol{\theta}}$. For discrete random variables the likelihood may also be interpreted as the probability of the model given the data, but this interpretation breaks down for continous random variables.

Log-Likelihood function:

- $l_n(\boldsymbol{\theta}|x_1, \ldots, x_n) = \sum_{i=1}^{n} \log f(x_i|\boldsymbol{\theta})$

The log-likelihood is additive over the samples x_i

Maximum Likelihood Estimator:



$$\hat{\boldsymbol{\theta}}^{MLE} = \arg\max l_n(\boldsymbol{\theta}|x_1, \ldots, x_n)$$

3.1.2 Recipe for obtaining MLEs

1. Specify probabilistic model
2. Write down log-likelihood function $l_n(\boldsymbol{\theta}|x_1, \ldots, x_n)$
3. Maximise $l_n(\boldsymbol{\theta}|x_1, \ldots, x_n)$

To maximise l_n one usually uses the **score function** $S(\boldsymbol{\theta})$

$$S(\theta) = \frac{dl_n(\theta|x_1,\ldots,x_n)}{d\theta}$$ scalar parameter: first derivative of log-likelihood function

$$\mathbf{S}(\boldsymbol{\theta}) = \nabla l_n(\boldsymbol{\theta}|x_1, \ldots, x_n)$$ gradient if $\boldsymbol{\theta}$ is a vector (i.e. if there's more than one parameter)

A necessary (but not sufficient) condition for the MLE is that

$$\boldsymbol{S}(\hat{\boldsymbol{\theta}}_{ML}) = 0$$

To demonstrate that the log-likelihood function actually achieves a maximum at $\hat{\boldsymbol{\theta}}_{ML}$ you also need to check that the curvature at the MLE is negative.

In the case of a single parameter (univariate θ) this requires to check that the second derivative of the log-likelihood function is negative:

$$\frac{d^2 l_n(\hat{\theta}_{ML})}{d\theta^2} < 0$$

In the case of a parameter vector (multivariate $\boldsymbol{\theta}$) you need to compute the Hessian matrix (matrix of second order derivatives, https://en.wikipedia.org/wiki/Hessian_matrix) at the MLE:

$$\nabla^T \nabla l_n(\hat{\boldsymbol{\theta}}_{ML})$$

and check if this matrix is negative definite (i.e. all its eigenvalues must be negative).

For a revisit of the Hessian matrix and related quantities see the refresher part of the lecture notes!

As we will see in Section 4 in the (negative of) second order derivative of the log-likelihood function also plays an important role for assessing the uncertainty of the MLE.

3.1.3 Invariance property of the MLE

Note that the maximisation is a procedure that is invariant against coordinate transformations of the argument: Suppose $x_{\max} = \arg\max h(x)$ and $y = g(x)$ where g is an invertible function. Then $y_{\max} = \arg\max h(g^{-1}(y)) = g(x_{\max})$. The achieved maximum itself remains invariant: $h(x_{\max}) = h(g^{-1}(y_{\max}))$.

With regard to maximum likelihood estimation this implies the following **invariance property** of the MLE:

- Suppose that $\hat{\theta}_{ML}$ is the MLE of θ.
- We transform the parameter to $\theta^\star = g(\theta)$ where g is an invertible function.
- Then $g(\hat{\theta}_{ML}) = \hat{\theta}^\star$ is the MLE of $\theta^\star$.
- The value of the achieved maximum likelihood is the same in both cases.

The invariance property of MLE can be very useful in practise, because it may be easier to perform the maximisation required for finding the MLE in a particular coordinate system.

3.2 Examples of maximum likelihood estimation

3.2.1 Example 1: Estimation of a proportion

- "Coin tossing" follows Bernoulli model: $\Pr(x = \text{"head"}) = p$, $\Pr(x = \text{"tail"}) = 1 - p$
- Parameter: p: probability of "heads"
- We conduct n trials and observe k "heads" and $n - k$ tails
- k follows Binomial distribution with $\mathrm{E}(k) = np$ and $\mathrm{Var}(k) = np(1 - p)$

What is the MLE of p?

1. likelihood function: $L(p) = \binom{n}{k} p^k (1-p)^{(n-k)}$
2. log-likelihood function: $l_n(p) = k \log p + (n - k) \log(1 - p) + C$
3. Score function $S(p) = \frac{dl_n(p)}{dp} = \frac{k}{p} - \frac{n-k}{1-p}$
   $\mathrm{S}(\hat{p}_{ML}) = 0 \rightarrow \hat{p}_{ML} = \frac{k}{n} \Rightarrow$ the relative frequency is the maximum likelihood estimator!
4. As $\frac{dS(p)}{dp} = -\frac{k}{p^2} - \frac{n-k}{(1-p)^2} < 0$ the optimum found in previous step corresponds indeed to the maximum of the (log-)likelihood function.

Note the **constant term** C **in the log-likelihood function** $l_n(p)$ that collects all terms that do not depend on the argument p. Specifically in the above it contains the Binomial coefficient with $C = \log \binom{n}{k}$. After taking the first derivative with regard to p this term disappears in $S(p)$, thus C **is not relevant for finding the MLE** of p. **In the future we will often omit** C **from the log-likelihood function without further mention.**

3.2.2 Example 2: Exponential Distribution

$$x \sim Exp(\lambda) \text{ with rate parameter } \lambda > 0$$

$$\mathrm{E}(x) = \frac{1}{\lambda} \text{ and } \mathrm{Var}(x) = \frac{1}{\lambda^2}$$

Density: $f(x) = \lambda \exp^{-\lambda x}$ with $\lambda > 0$

Data: $x_1, \ldots, x_n$

Likelihood function: $L(\lambda | x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i) = \lambda^n \exp(-\lambda \sum_{i=1}^{n} x_i)$

Log-Likelihood function: $l_n(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i = n(\log \lambda - \lambda \bar{x})$

Score function: $\frac{dl_n(\lambda)}{d\lambda} = S(\lambda) = n(\frac{1}{\lambda} - \bar{x})$

$$S(\hat{\lambda}_{ML}) = 0 \quad \Rightarrow \frac{1}{\hat{\lambda}_{ML}} = \bar{x} \Rightarrow \hat{\lambda}_{ML} = \frac{1}{\bar{x}}$$

Since $\frac{dS(\lambda)}{\lambda} = -\frac{n}{\lambda^2} < 0$ the optimum corresponds indeed to the maximum.

Note that the maximum likelihood estimator in this case is identical to the estimate obtained by the methods of moments (i.e. estimate mean by the average and then substitute in the population mean and solve for λ).

3.2.3 Example 3: Normal distribution with unknown mean and known variance

$$x \sim N(\mu, \sigma^2)$$
$$\mathrm{E}(x) = \mu$$
$$\mathrm{Var}(x) = \sigma^2$$

What's the maximum likelihood estimator for the parameter μ when σ^2 is known?

- Density:
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$
- Log-Density:
$$\log f(x) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}$$
- Log-likelihood function:
$$l_n(\mu) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \qquad \underbrace{-\frac{n}{2}\log(2\pi\sigma^2)}_{\text{constant, can be removed, does not depend on } \mu}$$

- Score function:
$$S(\mu) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)$$
- Maximum likelihood estimate:
$$S(\hat{\mu}_{ML}) = 0 \Rightarrow \hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$
- With $\frac{dS(\mu)}{d\mu} = -\frac{n}{\sigma^2} < 0$ the optimum is indeed the maximum
- Log-likelihood at maximum:
$$l_n(\hat{\mu}_{ML}) = -\frac{n}{2}\frac{s^2_{ML}}{\sigma^2} + \text{constant}$$
with $s^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ — this is the ML estimate of variance, see next examples.

Note the form of the log-likelihood function! Maximising $l_n(\mu) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$ is equivalent to *minimising* $\sum_{i=1}^{n}(x_i - \mu)^2$.

Hence, finding estimates by **maximum likelihood assuming a normal model is equivalent to least-squares estimation**. Note least-squares estimation is popular in regression and will be discussed later in the module.

3.2.4 Example 4: Normal Distribution with both mean and variance unknown

What's the maximum likelihood estimator for the parameter vector $\boldsymbol{\theta} = (\mu, \sigma^2)^T$?

- Density:
$$f(x) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$
- Log-likelihood function:
$$l_n(\boldsymbol{\theta}) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \qquad \underbrace{-\frac{n}{2}\log(2\pi)}_{\text{constant not depending on } \mu \text{ or } \sigma^2}$$
- Score function $\boldsymbol{S}$ (vector!), gradient of $l_n(\boldsymbol{\theta})$:
$$\boldsymbol{S}(\boldsymbol{\theta}) = \nabla l_n(\boldsymbol{\theta}) = \begin{pmatrix} \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 \end{pmatrix} = \begin{pmatrix} \frac{n}{\sigma^2}(\bar{x} - \mu) \\ -\frac{n}{2\sigma^2} + \frac{n}{2\sigma^4}\left(\overline{x^2} - 2\bar{x}\mu + \mu^2\right) \end{pmatrix}$$
with $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$.

Note that to obtain the second component of the score function the partial derivative needs to be taken with regard to the variance parameter σ^2 — not with regard to σ! Hint: replace $\sigma^2 = v$ in the log-likelihood function, then take the partial derivative with regard to v, then backsubstitute $v = \sigma^2$ in the result.

$$\boldsymbol{S}(\hat{\boldsymbol{\theta}}_{ML}) = 0 \Rightarrow \hat{\boldsymbol{\theta}}_{ML} = \begin{pmatrix} \hat{\mu}_{ML} \\ \widehat{\sigma^2}_{ML} \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} x_i \\ \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \end{pmatrix} = \begin{pmatrix} \bar{x} \\ \overline{x^2} - \bar{x}^2 \end{pmatrix}$$

- We look at the Hessian matrix in next section! Both eigenvalues are negative as required for a maximum!
- Log-likelihood at maximum:

$$l_n(\hat{\boldsymbol{\theta}}_{ML}) = -\frac{n}{2}\log(s^2_{ML}) - \frac{n}{2} + \text{constant}$$

Results from normal model:

1. The MLE for μ is the average:

$$\hat{\mu}_{ML} = m_{ML} = \bar{x}$$

2. The MLE of σ^2 is

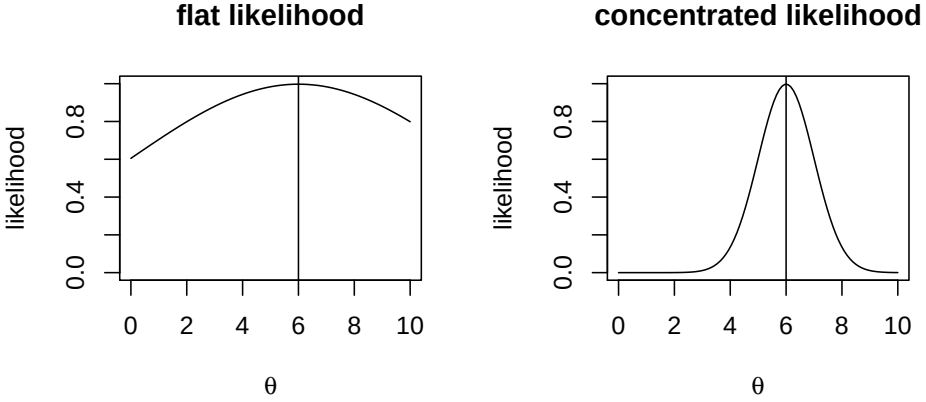
$$\widehat{\sigma^2}_{ML} = s^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Note the factor $\frac{1}{n}$ in front of the sum. This implies that $\widehat{\sigma^2}_{ML}$ is **biased** — recall that $\widehat{\sigma^2}_{UB} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is unbiased $\Rightarrow$ **Maximum likelihood can give biased estimates!**

3.3 Observed Fisher information

3.3.1 Motivation

This chapter explores the curvature the log-likelihood function, introduces the *observed* Fisher information and establishes the link to the *expected* Fisher information.



By inspection of some log-likelihood curves it is apparent that the log-likelihood function contains more information about the parameter $\boldsymbol{\theta}$ than just the maximum point $\hat{\boldsymbol{\theta}}_{ML}$.

In particular the **curvature** of the log-likelihood function must be somehow related the accuracy of $\hat{\boldsymbol{\theta}}_{ML}$: if the likelihood surface is flat near the maximum (low curvature) then if is more difficult to find the optimal parameter (also numerically!). Conversely, if the likelihood surface is peaked (strong curvature) then the maximum point is clearly defined.

3.3.2 Curvature of log-likelihood function

The curvature is linked to the second-order derivative of the log-likelihood function.

- Log-likelihood function:
$$l_n(\boldsymbol{\theta})$$

- First derivative of log-likelihood function = Score function:

  For univariate θ the score function is a scalar:
$$S(\theta) = \frac{dl_n(\theta)}{d\theta}$$

  For multivariate $\boldsymbol{\theta}$ of dimension d the score function is a vector of dimension d:
$$\boldsymbol{S}(\boldsymbol{\theta}) = \nabla l_n(\boldsymbol{\theta})$$

- Second derivative (=Hessian) of log-likelihood function:

  For univariate θ the Hessian is a scalar:
$$\frac{d^2 l_n(\theta)}{d\theta^2}$$

  For multivariate $\boldsymbol{\theta}$ the Hessian is a matrix of size $d \times d$:
$$\nabla^T \nabla l_n(\boldsymbol{\theta})$$

3.3.3 Observed Fisher information

Assume that the log-likelihood function $l_n(\boldsymbol{\theta})$ has a peak at the MLE $\hat{\boldsymbol{\theta}}_{ML}$. Then the curvature at the peak is by construction negative, i..e the Hessian matrix is negative definite at $\hat{\boldsymbol{\theta}}_{ML}$.

The **observed Fisher information** (matrix) is defined as the negative curvature at the MLE $\hat{\boldsymbol{\theta}}_{ML}$:
$$\boldsymbol{J}_n(\hat{\boldsymbol{\theta}}_{ML}) = -\nabla^T \nabla l_n(\hat{\boldsymbol{\theta}}_{ML})$$

Sometimes this is simply called the "observed information".

To avoid confusion with the expected Fisher information it is necessary to always use the qualifier "observed".

3.4 Observed Fisher information - Examples

3.4.1 Example 1: Bernoulli / Binomial model

We continue the Example 1 from Chapter 3. Recall that $\hat{p}_{ML} = \frac{k}{n}$ and the score function $S(p) = \frac{k}{p} - \frac{n-k}{1-p}$.

The negative second derivative of the log-likelihood function is:

$$-\frac{dS(p)}{dp} = \frac{k}{p^2} + \frac{n-k}{(1-p)^2}$$

The observed Fisher information is therefore

$$\begin{aligned} J_n(\hat{p}_{ML}) &= \frac{k}{\hat{p}_{ML}^2} + \frac{n-k}{(1-\hat{p}_{ML})^2} \\ &= \frac{k}{(k/n)^2} + \frac{n-k}{(1-(k/n))^2} \\ &= n(\frac{n}{k} + \frac{n}{n-k}) \\ &= n(\frac{1}{\hat{p}_{ML}} + \frac{1}{1-\hat{p}_{ML}}) \\ &= \frac{n}{\hat{p}_{ML}(1-\hat{p}_{ML})} \end{aligned}$$

The inverse of the observed Fisher information is:

$$J_n(\hat{p}_{ML})^{-1} = \frac{\hat{p}_{ML}(1-\hat{p}_{ML})}{n}$$

Compare with $\text{Var}\left(\frac{k}{n}\right) = \frac{p(1-p)}{n}$ (cf. Binomial distribution).

3.4.2 Example 2: Normal distribution

This is the continuation of Example 4 of Chapter 3. Recall the MLE for the mean and variance:

$$\hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

$$\widehat{\sigma^2}_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \overline{x^2} - \bar{x}^2$$

with $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$ and the score function (a vector of dimension 2):

$$\boldsymbol{S}_n(\mu, \sigma^2) = \nabla l_n(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2}(\bar{x} - \mu) \\ -\frac{n}{2\sigma^2} + \frac{n}{2\sigma^4}\left(\overline{x^2} - 2\mu\bar{x} + \mu^2\right) \end{pmatrix}$$

The Hessian matrix of the log-likelihood function is

$$\nabla^T \nabla l_n(\mu, \sigma^2) = \begin{pmatrix} -\frac{n}{\sigma^2} & -\frac{n}{\sigma^4}(\bar{x} - \mu) \\ -\frac{n}{\sigma^4}(\bar{x} - \mu) & \frac{n}{2\sigma^4} - \frac{n}{\sigma^6}\left(\overline{x^2} - 2\mu\bar{x} + \mu^2\right) \end{pmatrix}$$

The negative Hessian at the MLE, i.e. at $\hat{\mu}_{ML} = \bar{x}$ and $\widehat{\sigma^2}_{ML} = \overline{x^2} - \bar{x}^2$ yields the **observed Fisher information matrix**:

$$\boldsymbol{J}_n(\hat{\mu}_{ML}, \widehat{\sigma^2}_{ML}) = \begin{pmatrix} \frac{n}{\widehat{\sigma^2}_{ML}} & 0 \\ 0 & \frac{n}{2(\widehat{\sigma^2}_{ML})^2} \end{pmatrix}$$

Note that the observed Fisher information matrix is diagonal with positive entries. Therefore its eigenvalues are all positive as required for a maximum, because for a diagonal matrix the eigenvalues are simply the the entries on the diagonal.

The inverse of the observed Fisher information matrix is

$$\boldsymbol{J}_n(\hat{\mu}_{ML}, \widehat{\sigma^2}_{ML})^{-1} = \begin{pmatrix} \frac{\widehat{\sigma^2}_{ML}}{n} & 0 \\ 0 & \frac{2(\widehat{\sigma^2}_{ML})^2}{n} \end{pmatrix}$$

Recall that $x \sim N(\mu, \sigma^2)$ and therefore

$$\hat{\mu}_{ML} = \bar{x} \sim N(\mu, \frac{\sigma^2}{n})$$

Hence $\text{Var}(\hat{\mu}_{ML}) = \frac{\sigma^2}{n}$. If you compare this with the first entry of the inverse observed Fisher information matrix you see that this is essentially the same expression (apart from the "hat").

The empirical variance $\widehat{\sigma^2}_{ML}$ follows a scaled chi-squared distribution

$$\widehat{\sigma^2}_{ML} \sim \frac{\sigma^2}{n} \chi^2_{n-1}$$

with the following mean and variance:

- $\text{E}(\widehat{\sigma^2}_{ML}) = \frac{n-1}{n}\sigma^2$. This implies that $\widehat{\sigma^2}_{ML}$ is biased for small n, with $\text{Bias}(\widehat{\sigma^2}_{ML}) = \text{E}(\widehat{\sigma^2}_{ML}) - \sigma^2 = -\frac{1}{n}\sigma^2$. For large n we have $\text{E}(\widehat{\sigma^2}_{ML}) \stackrel{a}{=} \sigma^2$, so it becomes unbiased for large n.
- $\text{Var}(\widehat{\sigma^2}_{ML}) = \frac{n-1}{n}\frac{2\sigma^4}{n}$. For large n this becomes $\text{Var}(\widehat{\sigma^2}_{ML}) \stackrel{a}{=} \frac{2\sigma^4}{n}$ which is essentially (apart from the "hat") the second entry of the inverse observed Fisher information matrix!

3.4.3 Differences of observed to expected Fisher information

The observed and the expected Fisher information are superficially similar but in fact they are two quite different entities:

- Both types of Fisher information are based on computing the second order derivative (Hessian matrix).
- However, the observed Fisher informatio is computed from the log-likelihood function. Therefore it takes the observed data into account. It explicitly depends on the sample size n. It contains estimates of the parameters such as $\widehat{\sigma^2}$, not the parameters themselves. While the curvature

of the log-likelihood function can of course be computed for any point the the observed Fisher information refers to the curvature at the MLE. It is linked to the (asymptotic) variance of the MLE.

- In contrast, the expected Fisher information is derived directly from the log-density. It does not depend on observed data, and thus does not have dependency on sample size. It can be computed for any value of the parameters. It describes the geometry of the space of the models.

Chapter 4

Quadratic approximation and normal asymptotics

In this chapter we first introduce the multivariate normal distribution and then study second order approximation of the likelihood function.

4.1 Covariance, correlation and multivariate normal distribution

The density of a normally distributed scalar variable $x \sim N(\mu, \sigma^2)$ with mean $\mathrm{E}(x) = \mu$ and variance $\mathrm{Var}(x) = \sigma^2$ is

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

This is known as the univariate normal density. Note that the variance is given by

$$\mathrm{Var}(x) = \mathrm{E}\left((x - \mathrm{E}(x))^2\right) = \mathrm{E}\left((x-\mu)^2\right) = \mathrm{E}\left((x-\mu)(x-\mu)\right) = \mathrm{E}(x^2) - \mu^2$$

For a random vector $\boldsymbol{x} = (x_1, x_2, ..., x_d)^T$ with mean $\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\mu}$ the variance is generalised to the **covariance matrix** (of size $d \times d$).

$$\mathrm{Var}(\boldsymbol{x}) = \mathrm{E}\left(\underbrace{(\boldsymbol{x}-\boldsymbol{\mu})}_{d\times 1}\underbrace{(\boldsymbol{x}-\boldsymbol{\mu})^T}_{1\times d}\right) = \underbrace{\boldsymbol{\Sigma}}_{d\times d} = \mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T) - \boldsymbol{\mu}\boldsymbol{\mu}^T$$

The covariance matrix is symmetric by construction and **positive semi-definite**, i.e. the eigenvalues of $\boldsymbol{\Sigma}$ are all positive or equal to zero. However, we will aim to use non-singular covariance matrices, with all eigenvalues positive, so that it can be inverted.

A covariance matrix can factorised into the product

$$\boldsymbol{\Sigma} = \boldsymbol{V}^{\frac{1}{2}}\boldsymbol{P}\boldsymbol{V}^{\frac{1}{2}}$$

where $\boldsymbol{V}$ is a diagonal matrix containing the variances

$$\boldsymbol{V} = \begin{pmatrix} \sigma_{11} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{dd} \end{pmatrix}$$

and the matrix $\boldsymbol{P}$ ("capital rho") is the symmetric **correlation matrix**

$$\boldsymbol{P} = (\rho_{ij}) = \begin{pmatrix} 1 & \dots & \rho_{1d} \\ \vdots & \ddots & \vdots \\ \rho_{d1} & \dots & 1 \end{pmatrix}$$

Thus, the correlation between x_i and x_j is defined as

$$\rho_{ij} = \mathrm{Cor}(x_i, x_j) = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

The generalisation of the normal distribution to random vectors of size d is the **multivariate normal distribution** $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The corresponding density is

$$f(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{d}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \underbrace{\underbrace{(\boldsymbol{x}-\boldsymbol{\mu})^T}_{1\times d}\underbrace{\boldsymbol{\Sigma}^{-1}}_{d\times d}\underbrace{(\boldsymbol{x}-\boldsymbol{\mu})}_{d\times 1}}_{1\times 1=\text{scalar!}} \right)$$

For $d = 1$ we get $\boldsymbol{\mu} = \mu$ and $\boldsymbol{\Sigma} = \sigma^2$ and the multivariate normal density reduces to the univariate normal density.

4.2 Maximum likelihood estimates of the parameters of the multivariate normal distribution

Maximising the log-likelihood based on the multivariate normal density yields the MLEs for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. These are generalisations of the MLEs for the mean μ and variance σ^2 of the univariate normal as encounteed the previous chapter.

The estimates can be written in three different ways:

Data vector notation

with $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ the n vector-valued observations from the multivariate normal:

MLE for the mean:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{x}_k$$

MLE for the covariance:

$$\underbrace{\widehat{\boldsymbol{\Sigma}}}_{d\times d} = \frac{1}{n}\sum_{k=1}^{n} \underbrace{(\boldsymbol{x}_k - \hat{\boldsymbol{\mu}})}_{d\times 1} \underbrace{(\boldsymbol{x}_k - \hat{\boldsymbol{\mu}})^T}_{1\times d}$$

Note the factor $\frac{1}{n}$ in the estimator of the covariance matrix.

Data component notation

with x_{ki} the i-th component of the k-th sample:

$$\hat{\mu}_i = \frac{1}{n}\sum_{k=1}^{n} x_{ki}$$

$$\hat{\sigma}_{ij} = \frac{1}{n}\sum_{k=1}^{n} (x_{ki} - \hat{\mu}_i)\left(x_{kj} - \hat{\mu}_j\right)$$

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_d \end{pmatrix}, \widehat{\boldsymbol{\Sigma}} = (\hat{\sigma}_{ij})$$

Variance estimate:

$$\hat{\sigma}_{ii} = \frac{1}{n}\sum_{k=1}^{n} (x_{ki} - \hat{\mu}_i)^2$$

Data matrix notation

with $\boldsymbol{X} = (x_{ij})$ being the data matrix, with samples in rows and variables in columns.

Note that this is the *statistics convention* for the data matrix $\boldsymbol{X}$. However, *in the machine learning literature the convention is often reversed* and variables are assumed to be in the rows and samples in columns!!

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\boldsymbol{X}^T\mathbf{1}_n$$

Here $\mathbf{1}_n$ is a vector of length n containing 1 at each component.

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T$$

To simplify the expression for the estimate of the covariance matrix one often assumes that the data matrix is centered, i.e. that $\hat{\boldsymbol{\mu}} = 0$.

In *machine learning notation*, the data matrix has to be transposed and estimates are

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\boldsymbol{X}\mathbf{1}_n$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^T - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T$$

Because of the ambiguity in convention (machine learning vs statistics convention) and the often implicit use of centered data matrices the matrix notation is often confusing. Hence, using the other two notations is generally preferable.

4.3 Quadratic approximation of log-likelihood function around MLE

The observed Fisher information (matrix) $\boldsymbol{J}_n(\hat{\boldsymbol{\theta}}_{ML})$ occurs naturally in the quadratic approximation of the log-likelihood function:

Recall the Taylor series approximation of scalar-valued function $f(\boldsymbol{x})$ around $\boldsymbol{x}_0$:

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^T \nabla^T \nabla f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) + \dots$$

The second order Taylor series of $l_n(\boldsymbol{\theta})$ around the maximum likelihood $\hat{\boldsymbol{\theta}}_{ML}$ yields:

- for univariate θ:

$$l_n(\theta) = l_n(\hat{\theta}_{ML}) - \frac{1}{2}(\hat{\theta}_{ML} - \theta)^2 J_n(\hat{\theta}_{ML}) + \dots$$

- for multivariate $\boldsymbol{\theta}$:

$$l_n(\boldsymbol{\theta}) = l_n(\hat{\boldsymbol{\theta}}_{ML}) - \frac{1}{2}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta})^T \boldsymbol{J}_n(\hat{\boldsymbol{\theta}}_{ML})(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}) + \dots$$

Note that *there is no linear term* as we assume $\nabla l_n(\hat{\boldsymbol{\theta}}_{ML}) = 0$, i.e. that the gradient of the log-likelihood function vanishes at the MLE by construction.

Note the similarity of the above quadratic approximation with the log-density of the univariate and multivariate normal distribution:

- log-density of univariate normal distribution: $C - \frac{1}{2}(x - \mu)^2 \sigma^{-2}$
- log-density of multivariate normal: $C - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$

In particular note that in this approximation the observed Fisher information (matrix) plays the role of the **inverse** (co)variance (matrix) σ^{-2} and $\boldsymbol{\Sigma}^{-1}$!

Taking an quadratic approximation is thus closely linked to assuming normality.

4.4 Asymptotic normality of MLE

Theorem: Asymptotic normality of MLE point estimate, with inverse Fisher information as variance.

Intuitively, it would makes sense to associate large amount of curvature at the MLE with low variance of the MSE (and conversely, low amount of curvature with high variance).

This intuition is confirmed by the following theorem: **For large sample size n the MLE is normally distributed around the true parameter and with (co)variance equal to the inverse of the observed Fisher information**

$$\hat{\boldsymbol{\theta}}_{ML} \overset{a}{\sim} \underbrace{N_d}_{\text{multivariate normal}} \left(\underbrace{\boldsymbol{\theta}}_{\text{mean vector}} , \underbrace{\boldsymbol{J}_n(\hat{\boldsymbol{\theta}}_{ML})^{-1}}_{\text{covariance matrix}} \right)$$

For an single scalar parameter θ this reduces to

$$\hat{\theta} \overset{a}{\sim} N(\theta, J_n(\hat{\theta})^{-1})$$

This theorem is **valid under regularity conditions**. The most important requirements are that the likelihood is twice differentiable at $\hat{\boldsymbol{\theta}}_{ML}$ so that the observed Fisher information can be computed, and that the MLE lies at a peak within the support and not at the boundary.

Note we only state the result here, the proof itself will be explained in more advanced later modules (cf. Year 3 course "Statistical Inference"). Essentially, the proof works by showing that the error of the quadratic approximation becomes negligible for large sample size.

This theorem greatly enhances the usefulness of the method of maximum likelihood: In regular setting ML not only yields point estimates for the parameters but also (asymptotic) estimates of their variance and a corresponding normal sampling distribution for the estimated parameter.

4.5 Observed or expected Fisher information to estimate variance of the MLE?

There used to be some discussion whether to use the observed or the expected Fisher information to estimate the variance. There's an important classic paper that answers this question:

Efron, B. & Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. Biometrika, 65, 457-87. https://doi.org/10.1093/biomet/65.3.457

Conclusion: **use the observed** Fisher information $\boldsymbol{J}_n$ **not the expected** Fisher information $\boldsymbol{I}$!
Because the observed Fisher information is based on the data at hand (like the MLE).

4.6 Normal confidence intervals for MLEs

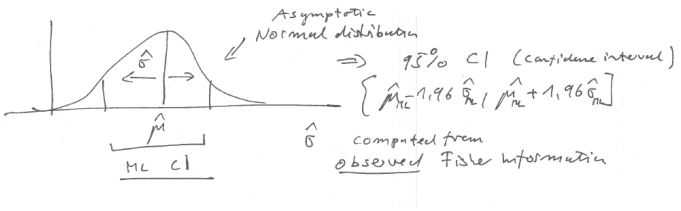
Given a probability model for the data, maximum likelihood proceeds by

- maximisation of $l_n(\boldsymbol{\theta})$
- computation of the curvature at $l_n(\hat{\boldsymbol{\theta}}^{ML})$.

Then you get:

- a point estimate $\hat{\boldsymbol{\theta}}_{ML}$
- the asymptotic variance of $\hat{\boldsymbol{\theta}}_{ML}$
- the corresponding asymptotic normal distribution

The asymptotic normality enables us to construct a corresponding normal confidence interval (CI) and also to conduct associated tests (is a value included in the CI or not?):



Thus, to construct the asymptotic normal CI for a maximum likelihood estimator of a scalar θ we use the MLE $\hat{\theta}_{ML}$ and its standard deviation $\hat{\sigma} = \widehat{\mathrm{SD}}(\hat{\theta}_{ML})$ computed from the observed Fisher information:

$$\mathrm{CI} = [\hat{\theta}_{ML} \pm c_{normal}\hat{\sigma}]$$

c_{normal} is a critical value for the standard-normal symmetric confidence interval chosen to achieve the desired nominal coverage (see also refresher).

κ coverage	Critical value c_{normal}
0.9	1.64
0.95	1.96
0.99	2.58

These values are computed using the inverse standard normal distribution function via $c_{\mathrm{normal}} = \Phi^{-1}\left(\frac{1+\kappa}{2}\right)$.

For example, for a CI with 95% coverage one uses the factor 1.96 so that

$$\mathrm{CI} = [\hat{\theta}_{ML} \pm 1.96\,\widehat{\mathrm{SD}}(\hat{\theta}_{ML})]$$

.

4.7 Wald statistic

Centering the MLE $\hat{\boldsymbol{\theta}}_{ML}$ with $\boldsymbol{\theta}_0$ followed by standardising with $\widehat{\text{SD}}(\hat{\theta}_{ML})$ yields the **Wald statistic**:

(for scalar θ)

$$t(\theta_0) = \frac{\hat{\theta}_{ML} - \theta_0}{\text{SD}(\hat{\theta}_{ML})} = J_n(\hat{\theta}_{ML})^{1/2}(\hat{\theta}_{ML} - \theta_0)$$

(for vector $\boldsymbol{\theta}$)

$$\boldsymbol{t}(\boldsymbol{\theta}_0) = \text{SD}(\hat{\boldsymbol{\theta}}_{ML})^{-1}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0) = \boldsymbol{J}_n(\hat{\boldsymbol{\theta}}_{ML})^{1/2}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0)$$

Note in the multivariate case we need to use *matrix inversion* and the *matrix square root*.

We now assume that the true underlying parameter is $\boldsymbol{\theta}_0$. Since the MLE is asymptotically normal the Wald statistic is asymptotically **standard normal** distributed as follows:

$$\begin{array}{lll} \boldsymbol{t}(\boldsymbol{\theta}_0) \overset{a}{\sim} & N_d(0, \boldsymbol{I}_d) & \text{for vector } \boldsymbol{\theta} \\ t(\theta_0) \overset{a}{\sim} & N(0,1) & \text{for scalar } \theta \end{array}$$

Correspondingly, the **squared** Wald statistic is chi-squared distributed assuming $\boldsymbol{\theta}_0$ as true parameter:

$$\begin{array}{lll} \boldsymbol{t}(\boldsymbol{\theta}_0)^T \boldsymbol{t}(\boldsymbol{\theta}_0) & \overset{a}{\sim} \chi^2_d & \text{for vector } \boldsymbol{\theta} \\ t(\theta_0)^2 & \overset{a}{\sim} \chi^2_1 & \text{for scalar } \theta \end{array}$$

4.8 Normal CI expressed using the squared Wald statistics

The normal CI can be expressed using Wald statistics as follows:

$$\text{CI} = \{\theta_0 : |t(\theta_0)| < c_{\text{normal}}\}$$

Similary, it can also be expressed using the squared Wald statistics:

$$\text{CI} = \{\theta_0 : t(\theta_0)^2 < c_{\text{chisq}}\}$$

The following list the critical values for the three most common choice of coverage κ for $df = 1$ when using the chi-squared distribution:

κ coverage	Critical value c_{chisq} ($df = 1$)
0.9	2.71
0.95	3.84
0.99	6.63

4.9 Testing and confidence intervals

There is a **duality between confidence intervals and statistical tests**: for every θ_0 inside a CI with coverage κ the data do not allow to reject the hypothesis that θ_0 is the true parameter with significance level $1-\kappa$. In contrast, all values θ_0 outside the CI can be rejected with significance level $1-\kappa$ to be the true parameter.
Thus, **the test decision (reject or not) is mirrored in CIs by whether a parameter lies outside or inside the CI.**

Therefore, the Wald statistic can be used both as a statistic to test whether $\boldsymbol{\theta}_0$ is the true underlying parameter value as well as to construct CIs covering the true parameter.

4.10 Example: normal distribution

$x_1, \ldots, x_n \sim N(\mu, \sigma^2)$ $\quad$ $\mu = ?$, $\sigma^2 =$ constant/known

$l_n(\mu) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$ $\quad$ log-likelihood

$S_n(\mu) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)$

$$\hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i) = \bar{x}$$

The corresponding observed Fisher information at $\hat{r}_{ML} = \hat{\mu}_{ML}$:

$$J_n(\hat{\mu}_{ML}) = \frac{n}{\sigma^2}$$

Asymptotic distribution of $\hat{\mu}_{ML}$:

$$\hat{\mu}_{ML} \overset{a}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

Note that in this case it is in fact also the **exact** solution (not just valid asymptotically).

Wald statistic:

$$t(\mu_0) = \frac{\hat{\mu}_{ML} - \mu_0}{\sigma/\sqrt{n}} \overset{a}{\sim} N(0,1)$$

This is the one sample t-statistic (with given σ).

Squared Wald statistic:

$$t(\mu_0)^2 = \frac{(\hat{\mu}_{ML} - \mu_0)^2}{\sigma^2/n} \overset{a}{\sim} \chi^2_1$$

Using the Wald or the squared Wald statistics we can test whether μ_0 can be rejected as underlying true parameter, and we can construct corresponding confidence intervals.

4.11 Example of non-regular model

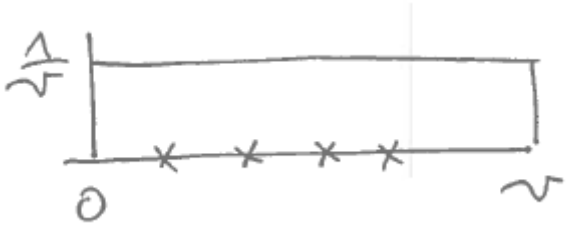
Example of a model with non-differentiable likelihood function at the MLE is the uniform distribution with upper bound θ:

$$x_1, \ldots, x_n \sim U(0, \theta)$$

$$\hat{\theta}_{ML} = ?$$

With $x_{[i]}$ we denote the *ordered* observations with $0 \leq x_{[1]} < x_{[2]} < \ldots < x_{[n]} \leq \theta$ and $x_{[n]} = \max(x_1, \ldots, x_n)$. The probability density function of $U(0, \theta)$ is

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & \text{if } x \in [0, \theta] \\ 0 & \text{otherwise} \end{cases}$$



and on log-scale

$$\log f(x|\theta) = \begin{cases} -\log \theta & \text{if } x \in [0, \theta] \\ -\infty & \text{otherwise} \end{cases}$$

Since all observed data $x_1, \ldots, x_n$ lie in the interval $[0, \theta]$ we get as log-likelihood function

$$l_n(\theta) = -n \log \theta$$

with the condition $x_{[n]} \leq \theta$. Therefore the log-likelihood function is maximised at $\hat{\theta}_{ML} = x_{[n]}$.

Note that $l_n(\theta)$ **is not differentiable** at $\hat{\theta}_{ML}$ because it sits at the border of the allowed range for θ. This means thta the **observed Fisher information cannot be computed** and the asymptotic **normal approximation is not available.**

Nonetheless, we can still obtain the sampling distribution of $\hat{\theta}_{ML} = x_{[n]}$. However, *not* via asymptotic ML arguments but instead by understanding that $x_{[n]}$ is an order statistic (see https://en.wikipedia.org/wiki/Order_statistic) with the following properties:

$$
\begin{aligned}
x_{[n]} &\sim \theta \,\mathrm{Beta}(n,1) && \text{"n-th order statistic"} \\
\mathrm{E}(x_{[n]}) &= \tfrac{n}{n+1}\theta \\
\mathrm{Var}(x_{[n]}) &= \tfrac{n}{(n+1)^2(n+2)}\theta^2 && \approx \tfrac{\theta^2}{n^2}
\end{aligned}
$$

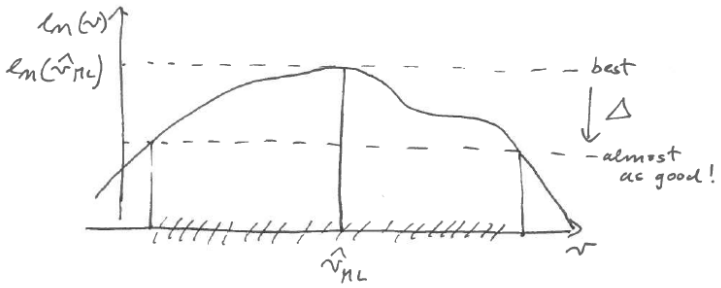
Note that the variance decreases with $\frac{1}{n^2}$ which is much faster than the usual $\frac{1}{n}$, which makes $\hat{\theta}_{ML}$ a super efficient estimator.

Chapter 5

Likelihood-based confidence interval and likelihood ratio

5.1 Likelihood-based confidence intervals

Instead of relying on normal / quadratic approximation, we can also use the log-likelihood directly to find the so called **likelihood confidence intervals**:



Idea: find all θ_0 that have a log-likelihood that is almost as good as $l_n(\hat{\theta}_{ML})$.

$$\text{CI} = \{\theta_0 : l_n(\hat{\theta}_{ML}) - l_n(\theta_0) \leq \Delta\}$$

Here Δ is the tolerated deviation from the maximum log-likelihood. We will see below how to determine a suitable Δ further below.

Advantages of using likelihood-based CI:

- not restricted to be symmetric
- enables to construct multivariate CIs for parameter vector easily even in non-normal cases
- contains normal CI as special case

Question: how to choose Δ, i.e how to calibrate the likelihood interval? Essentially, by comparing with normal CI!

5.2 Wilks log likelihood ratio statistic

The **Wilks likelihood ratio statistic** W is defined as:

$$W(\boldsymbol{\theta}_0) = 2\log\left(\frac{L(\hat{\boldsymbol{\theta}}_{ML})}{L(\boldsymbol{\theta}_0)}\right) = 2(l_n(\hat{\boldsymbol{\theta}}_{ML}) - l_n(\boldsymbol{\theta}_0))$$

Note we can write the likelihood CI in terms of Wilk's W as follows:

$$\text{CI} = \{\boldsymbol{\theta}_0 : W(\boldsymbol{\theta}_0) \leq 2\Delta\}$$

5.3 Quadratic approximation of Wilks statistic

Recall the *quadratic approximation* of the log-likelihood function $l_n(\boldsymbol{\theta})$ around the MLE $\hat{\boldsymbol{\theta}}_{ML}$:

$$l_n(\boldsymbol{\theta}) \approx l_n(\hat{\boldsymbol{\theta}}_{ML}) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{ML})^T \boldsymbol{J}_n(\hat{\boldsymbol{\theta}}_{ML})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{ML})$$

$$\Rightarrow W(\boldsymbol{\theta}_0) \approx (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{ML})^T \boldsymbol{J}_n(\hat{\boldsymbol{\theta}}_{ML})(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{ML}) = \boldsymbol{t}(\boldsymbol{\theta}_0)^T \boldsymbol{t}(\boldsymbol{\theta}_0)$$

Thus the quadratic approximation of the Wilks statistics W yields the squared Wald statistic!

5.4 Distribution of Wilks statistics

The connection with the squared Wald statistics implies that both have asympotically the same distribution.

Thus, W is asymptotically χ^2_d distributed (with d degrees of freedom if $\boldsymbol{\theta}$ has dimension d).

For scalar θ (i.e. single parameter): $d = 1$ and $W \sim \chi^2_1$

5.4.1 Cutoff values Δ

The asymptotic distribution for W is useful to choose a suitable Δ for the likelihood CI since $2\Delta = c_{chisq}$.

This yields for scalar θ ($df = 1$):

κ coverage	Δ ($df = 1$)
0.9	1.35
0.95	1.92
0.99	3.32

5.5 Example: likelihood CI for exponential model

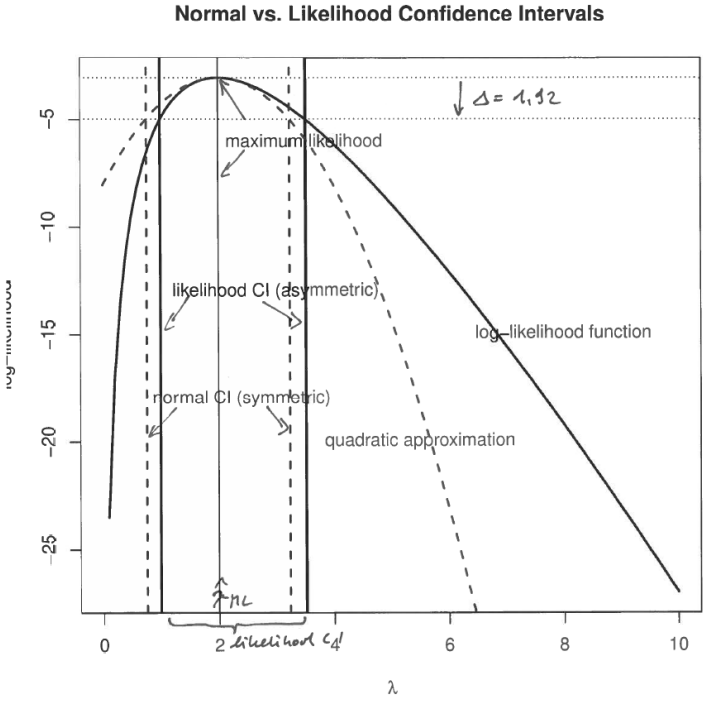
We consider the parameter λ in an exponential distribution (cf. Question 4 in Worksheet 2).

We observe $n = 10$ observations $x_1, \ldots, x_n$ with average $\frac{1}{10}\sum_{i=1}^{10} x_i = \bar{x} = 1/2$.

The corresponding log-likelihood function for this data is $l_{10}(\lambda) = n \log \lambda - \lambda n \bar{x} = 10 \log \lambda - 5\lambda$.

From analytic calculations (cf. Worksheet 2) we know that the maximum likelihood estimate is $\hat{\lambda}_{ML} = \frac{1}{\bar{x}} = 2$ and that the observed Fisher information is $J_{10}(\hat{\lambda}_{ML}) = n\bar{x}^2 = 5/2$. Thus, the estimated asymptotic standard deviation is $\mathrm{SD}(\hat{\lambda}_{ML}) = \sqrt{2/5} \approx 0.63$, so that the asymptotic normal symmetric confidence interval with 95% coverage for λ is $[2 \pm 1.24]$.

The following figure shows the Likelihood CI and the normal CI for the parameter λ in the exponential model



In this example note that

- the Likelihood CI is shifted when compared to a normal CI.
- The normal CI is identical to the likelihood CI using the quadratic approximation

5.6 Origin of likelihood ratio statistic

Assume that F is the true (and unknown) data generating model and we would like to compare two candidate models G_A and G_B on the basis of observed data $x_1, \ldots, x_n$. The KL divergences $D_A = D_{\mathrm{KL}}(F, G_A)$ and $D_B = D_{\mathrm{KL}}(F, G_B)$ indicate how close each of the models G_A and G_B fit the true F. The difference $D_B - D_A$ is thus a way to measure the relative fit of the two models, and can be computed as

$$D_B - D_A = D_{\mathrm{KL}}(F, G_B) - D_{\mathrm{KL}}(F, G_A) = \mathrm{E}_F \log \frac{f_A(x)}{f_B(x)}$$

Replacing F by the empirical distribution $\hat{F}_n$ leads to the large sample approximation

$$D_B - D_A \approx \frac{1}{n}(l_n(\theta_A) - l_n(\theta_B))$$

Hence, the difference in the log-likelihoods provides an estimate of the difference in the KL divergence of the two models involved.

The Wilks likelihood ratio statistic

$$W(\boldsymbol{\theta}_0) = 2 \log \left(\frac{L(\hat{\boldsymbol{\theta}}_{ML})}{L(\boldsymbol{\theta}_0)} \right) = 2(l_n(\hat{\boldsymbol{\theta}}_{ML}) - l_n(\boldsymbol{\theta}_0))$$

thus compares the best-fit distribution with $\hat{\boldsymbol{\theta}}_{ML}$ as the parameter to the distribution with parameter $\boldsymbol{\theta}_0$.

5.7 Distribution of Wilks statistic and Likelihood CI

Under $\boldsymbol{\theta}_0$ the Wilks statistic is distributed asymptotically as

$$W(\boldsymbol{\theta}_0) \overset{a}{\sim} \chi^2_d$$

where $|\boldsymbol{\theta}| = d$ is the number of parameters, i.e. dimension of the model. Note that in the above only the numerator is optimised (to find the MLE $\hat{\boldsymbol{\theta}}_{ML}$), the denominator is fixed at a specified $\boldsymbol{\theta}_0$.

As discussed earlier the Wilks statistic can be used to obtain a confidence interval which includes all models with parameter $\boldsymbol{\theta}_0$ that are not much worse in terms of likelihood than the best one ($\hat{\boldsymbol{\theta}}_{ML}$):

$$CI = \{\boldsymbol{\theta}_0 : W(\boldsymbol{\theta}_0) \leq c\}$$

The critical value c is obtained from from the χ^2_d distribution, e.g. $c = 3.84$ if $d = 1$ and $\kappa = 0.95$ coverage is desired.

5.8 Likelihood ratio test (LRT)

$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ True model is $\boldsymbol{\theta}_0$ $\rightarrow$ Null hypothesis / null model = simple
$H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ True model is **not** $\boldsymbol{\theta}_0$ $\rightarrow$ Alternative hypothesis / alternative models = composite

In order to test whether H_0 is true we need to find a suitable test statistic. Extreme values of this test statistic imply evidence against H_0. In a likelihood ratio test the test statistic is chosen to be

$$W(\boldsymbol{\theta}_0) = 2(l_n(\hat{\boldsymbol{\theta}}_{ML}) - l_n(\boldsymbol{\theta}_0))$$

or equivalently

$$\Lambda(\boldsymbol{\theta}_0) = \frac{L(\boldsymbol{\theta}_0)}{L(\hat{\boldsymbol{\theta}}_{ML})}$$

They can be transformed into each other by $W(\boldsymbol{\theta}_0) = -2\log\Lambda(\boldsymbol{\theta}_0)$ and $\Lambda(\boldsymbol{\theta}_0) = e^{-1/2W(\boldsymbol{\theta}_0)}$.

Remarks:

- The composite alternative H_1 is represented by a single point (the MLE).
- **Reject** H_0 for **large values of** W or equivalently for **small values of** Λ .
- Wilks' theorem: under H_0 and for large n the statistic W is chi-squared distributed, i.e. $W \overset{a}{\sim} \chi^2_d$. This allows to compute critical values (i.e tresholds to declared rejection under a given significance level) and also p-values corresponding to the observed test statistics.
- Models **outside** the CI are **rejected**
- Models **inside** the CI **cannot be rejected**, i.e. they can't be statistically distinguished from the best alternative model.

5.9 Optimality of LRTs

LRT statistics is asymptotically linked to differences in the KL divergences of the two compared model with the true model. It can be shown that the LRT statistic to compare two simple model is optimal in the sense that for any given specified type I error (=probability of wrongly rejecting H_0, i.e. the sigificance level) it will maximise power (=1- type II error, probability of correctly accepting H_1). This is know as the Neyman-Pearson theorem (more details to come in year 3!).

As a result, the likelihood framework not only allows to find asymptotically optimal point and interval estimators but also allows optimal inference (testing).

5.10 Generalised likelihood ratio test (GLRT)

Also known as **maximum likelihood ratio test (MLRT)**. The Generalised Likelihood Ratio Test (GLRT) works like the previous test with the difference that

now the null model H_0 is composite as well. This means that in the denominator in the test statistics needs to be optimised as well.

$H_0 : \boldsymbol{\theta} \in \omega_0 \subset \Omega$	True model lies in restricted model space
$H_1 : \boldsymbol{\theta} \in \omega_1 = \Omega \setminus \omega_0$	True model is not the restricted model space

Both H_0 and H_1 are now composite hypotheses. Ω represents the unrestricted model space with dimension (=number of free parameters) $d = |\Omega|$. The constrained space ω_0 has degree of freedom $d_0 = |\omega_0|$ with $d_0 < d$. Note that in the standard LRT the set ω_0 is a simple point with $d_0 = 0$ as the null model is a simple distribution. Thus, LRT is contained in GLRT as special case!

The corresponding generalised likelihood ratio statistics are given by

$$W = 2\log\left(\frac{L(\hat{\theta}_{ML})}{L(\hat{\theta}^0_{ML})}\right) \text{ and } \Lambda = \frac{\max\limits_{\theta\in\omega_0} L(\theta)}{\max\limits_{\theta\in\Omega} L(\theta)}$$

where $L(\hat{\theta}_{ML})$ is the maximised likelihood assuming the full model (with parameter space Ω) and $L(\hat{\theta}^0_{ML})$ is the maximised likelihood for the restricted model (with parameter space ω_0).

Remarks:

- MLE in the restricted model space ω_0 is taken as a representative of H_0.
- The likelihood is **maximised** in **both numerator** and **denominator**.
- The restriced model is a special case of the full model (i.e. the two models are nested).
- The asymptotic distribution of W is chi-squared with degree of freedom depending on both d and d_0:

$$W \overset{a}{\sim} \chi^2_{d-d_0}$$

- If H_0 is a simple hypothesis (i.e. $d_0 = 0$) then the standard LRT (and corresponding CI) is recovered as special case of the GLRT.

5.11 GLRT example

Case-control study: (e.g. "healthy" vs. "disease")
we observe normal data from two groups with sample size n_1 and n_2 (and $n = n_1 + n_2$):

$$x_1, \ldots, x_{n_1} \sim N(\mu_1, \sigma^2)$$

and

$$x_{n_1+1}, \ldots, x_n \sim N(\mu_2, \sigma^2)$$

Question: are the two means μ_1 and μ_2 the same in the two groups?

$H_0 : \mu_1 = \mu_2$ (with variance unknown nuisance parameter)
$H_1 : \mu_1 \neq \mu_2$

Restricted and full models:

ω_0: restricted model with two parameters μ_0 and σ_0^2 (so that $x_1, \ldots, x_n \sim N(\mu_0, \sigma_0^2)$).

Ω: full model with three parameters μ_1, μ_2, σ^2.

Corresponding log-likelihood functions:

Restricted model ω_0:

$$\log L(\mu_0, \sigma_0^2) = -\frac{n}{2}\log(\sigma_0^2) - \frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i - \mu_0)^2$$

Full model Ω:

$$\begin{aligned}\log L(\mu_1, \mu_2, \sigma^2) &= \left(-\frac{n_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n_1}(x_i - \mu_1)^2\right) + \left(-\frac{n_2}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=n_1+1}^{n}(x_i - \mu_1)^2\right) \\ &= -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n_1}(x_i - \mu_1)^2 + \sum_{i=n_1+1}^{n}(x_i - \mu_2)^2\right)\end{aligned}$$

Corresponding MLEs:

$$\begin{aligned}\omega_0 : \quad & \hat{\mu}_0 = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} x_i \qquad \widehat{\sigma_0^2} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(x_i - \hat{\mu}_0)^2 \\ \Omega : \quad & \hat{\mu}_1 = \tfrac{1}{n_1}\textstyle\sum_{i=1}^{n_1} x_i \qquad \widehat{\sigma^2} = \tfrac{1}{n}\left\{\textstyle\sum_{i=1}^{n_1}(x_i - \hat{\mu}_1)^2 + \sum_{i=n_1+1}^{n}(x_i - \hat{\mu}_2)^2\right\} \\ & \hat{\mu}_2 = \tfrac{1}{n_2}\textstyle\sum_{i=n_1+1}^{n} x_i\end{aligned}$$

Corresponding maximised log-likelihood:

Restricted model:

$$\log L(\hat{\mu}_0, \widehat{\sigma_0^2}) = -\frac{n}{2}\log(\widehat{\sigma_0^2}) - \frac{n}{2}$$

Full model:

$$\log L(\hat{\mu}_1, \hat{\mu}_2, \widehat{\sigma^2}) = -\frac{n}{2}\log(\widehat{\sigma^2}) - \frac{n}{2}$$

Likelihood ratio statistic:

$$W = 2\log\left(\frac{L(\hat{\mu}_1, \hat{\mu}_2, \widehat{\sigma^2})}{L(\hat{\mu}_0, \widehat{\sigma_0^2})}\right) = 2\log L(\hat{\mu}_1, \hat{\mu}_2, \widehat{\sigma^2}) - 2\log L(\hat{\mu}_0, \widehat{\sigma_0^2}) = n\log\left(\frac{\widehat{\sigma_0^2}}{\widehat{\sigma^2}}\right)$$

Using $\widehat{\sigma_0^2} - \widehat{\sigma^2} = \frac{n_1 n_2}{n^2}(\hat{\mu}_1 - \hat{\mu}_2)^2$ this can be further simplified:

$$W = n \log\left(\frac{\widehat{\sigma_0^2}}{\widehat{\sigma^2}}\right) = n \log\left(1 + \frac{t_{ML}^2}{n}\right) = n \log\left(1 + \frac{1}{n-2}t^2\right)$$

with (ML variance)

$$t_{ML} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\widehat{\sigma^2}}}$$

and (unbiased variance)

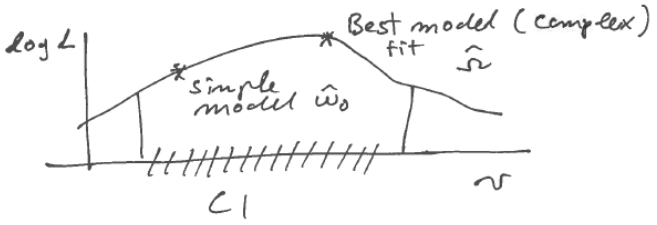
$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\frac{n}{n-2}\widehat{\sigma^2}}}$$

$\longrightarrow$ the GRLT is a monotone function of the (squared) two-sample t-statistic!

It can be shown that all standard tests with normal distributions can be interpreted as GLRTs!

5.12 Thoughts on model selection

- Note that, by construction, the model with more parameters always has a higher likelihood, implying likelihood favours complex models
- However, in both the LRT and GLRT in order to "win" the more complex model (based on optimising Ω) must be significantly better / have significantly higher likelihood than the simpler model (based on optimising ω_0)
- If we cannot reject the simpler model ω_0, then it will be included in the CI.



- Recall that the aim in statistics is **not** about rejecting models (this is easy as for large sample size any model will be rejected!)

- Instead, the aim is model building, i.e. to find a model that **explains the data well** and that **predicts well**!

- Typically, this will not be the best-fit ML model, but rather a simpler model that is close enough to the best / most complex model

- Complex model may overfit!

Chapter 6

Optimality properties, minimal sufficiency and summary

6.1 Properties of MLEs encountered so far

1. MLE is a special case of relative entropy minimisation *valid for large samples*.
2. MLE can be seen as generalisation of least squares (and conversely, least squares is a special case of ML).

Kullback-Leibler 1951
Entropy learning: minimise $D_{\mathrm{KL}}(F_{\mathrm{true}}, F_{\boldsymbol{\theta}})$
$\downarrow$
large n
$\downarrow$
Fisher 1922
Maximise Likelihood $L(\boldsymbol{\theta}|x_1, \ldots, x_n)$
$\downarrow$
normal model
$\downarrow$
Gauss 1822
Minimise squared error $\sum_i (x_i - \theta)^2$

3. Given a model, derivation of the MLE is basically automatic (only optimisation required)!
4. MLEs are *not* necessarily unbiased as in the example of the variance of the normal model (but MLEs are *asymptotically* unbiased as we will see later)
5. MLE's are invariant against parameter transformations $\rightarrow$ invariance property.

6. MLEs are **asympotically normally distributed**, with asymptotic variance determined by the inverse negative curvature that the MLE.

   Remark: there are methods to obtain higher-order (non-normal) asymptotic approximations for the distribution of $\hat{\boldsymbol{\theta}}_{ML}$. This is called higher order likelihood inference.

6.2 Further optimality properties of MLEs

For a very wide class of models and in *regular situations* (i.e. smooth and differentiable likelihood, maximum lies within parameter boundaries, number of parameters small compared to sample size. etc.) estimators constructed by maximum likelihood enjoy a number of highly favourable optimality properties.

The precise mathematical details of the regularity conditions are *not* subject of this course but will be discussed in year 3 and 4 modules in Statistics.

1. MLEs are **consistent**: if the true underlying model F_{true} with parameter $\boldsymbol{\theta}_{\text{true}}$ is contained in the set of specified candidates models $F_{\boldsymbol{\theta}}$

   $$\underbrace{F_{\text{true}}}_{\text{true model}} \subset \underbrace{F_{\boldsymbol{\theta}}}_{\text{specified models}}$$

   then

   $$\hat{\boldsymbol{\theta}}_{ML} \overset{\text{large } n}{\longrightarrow} \boldsymbol{\theta}_{\text{true}}$$

   This is a consequence of $D_{\text{KL}}(F_{\text{true}}, F_{\boldsymbol{\theta}}) \to 0$ for $F_{\boldsymbol{\theta}} \to F_{\text{true}}$. Correspondingly, **MLEs are asympotically unbiased** (but can be biased in finite samples!)

   Note that even if the candidate model $F_{\boldsymbol{\theta}}$ is misspecified (i.e. it does not contain the actual true model) the MLE is still optimal in the sense in that it will find the closest possible model.

   It is possible to construct inconsistent MLEs, but this occurs only in situations where the dimension of the model / number of parameters increases with sample size, or when the MLE is at a boundary or when there are singularities in the likelihood function

2. MLEs are **asympotically optimally efficient** (Cramer-Rao theorem): For large samples the MLE achieves the lowest possible variance possible in an estimator — this is the so-called Cramer-Rao lower bound. The variance decreases to zero with $n \to \infty$ typically with rate $1/n$.

Hence, for large sample size n the best estimator will typically be the MLE.

However, for **small sample size it is indeed possible (and necessary) to improve over the MLE** (e.g. via Bayesian estimation or regularisation).

6.3 Summarising data and the concept of minimal sufficiency

Closely linked with likelihood theory is the concept of a **minimally sufficient statistic** to optimally summarise the information available in the data about a parameter in a model.

Generally, a **statistic** $T(x_1, \ldots, x_n) = T(x_i)$ is function of the data $x_1, \ldots, x_n$. In the following we write x_i as a shorthand for the complete data set with n observations. The statistic $T(x_i)$ can be of any type and value (scalar, vector, matrix etc. — even a function). $T(x_i)$ is called a *summary statistic* if it describes important aspects of the data such as location (e.g. the average $\text{avg}(x_i) = \bar{x}$, the median) or scale (e.g. standard deviation, interquartile range).

A statistic $T(x_i)$ is said to be **sufficient** for a parameter $\boldsymbol{\theta}$ in a model if the corresponding likelihood function can be written in terms of $T(x_i)$ so that

$$L(\boldsymbol{\theta}|x_i) = h(T(x_i), \boldsymbol{\theta})\, k(x_i)\,,$$

where $h(x)$ and $k(x)$ are positive-valued functions, and or equivalently on log-scale

$$l_n(\boldsymbol{\theta}) = \log h(T(x_i), \boldsymbol{\theta}) + \log k(x_i)\,.$$

This is known as the **Fisher-Pearson factorisation**. By construction, estimation and inference about $\boldsymbol{\theta}$ based on the factorised likelihood $L(\boldsymbol{\theta})$ is mediated through the sufficient statistic $T(x_i)$ and does not require the original data x_i. Instead, the sufficient statistic $T(x_i)$ contains all the information in x_i required to learn about the parameter $\boldsymbol{\theta}$. Therefore, if the MLE $\hat{\boldsymbol{\theta}}_{ML}$ of $\boldsymbol{\theta}$ exists and is unique then **the MLE is a unique function of the sufficient statistic** $T(x_i)$. If the MLE is not unique then it can be chosen to be function of $T(x_i)$. Note that **a sufficient statistic always exists** since the data x_i are themselves sufficient statistics, with $T(x_i) = x_i$. Furthermore, sufficient statistics are **not unique** since applying a one-to-one transformation to $T(x_i)$ yields another sufficient statistic.

Every sufficient statistic $T(x_i)$ induces a partitioning of the space of data sets by clustering all hypothetical outcomes for which the statistic $T(x_i)$ assumes the same value t:

$$\mathcal{X}_t = \{x_i : T(x_i) = t\}$$

The **data sets in $\mathcal{X}_t$ are equivalent in terms of the sufficient statistic** $T(x_i)$. Note that the dimensions of $T(x_i)$ may be much smaller than those of x_i. Instead of n data points as few as one or two summaries may be sufficient to fully convey all the information in the data about the model parameters. Thus, transforming data x_i using a sufficient statistic $T(x_i)$ may result in substantial **data reduction**.

Data sets x_i and y_i for which the ratio of the likelihoods $L(\boldsymbol{\theta}|x_i)/L(\boldsymbol{\theta}|y_i)$ does not depend on $\boldsymbol{\theta}$ (so the two likelihoods are proportional to each other by a constant) are called **likelihood equivalent** because a likelihood-based procedure to learn about $\boldsymbol{\theta}$ will draw identical conclusions from x_i and y_i. For data sets $x_i, y_i \in \mathcal{X}_t$ equivalent with respect to a sufficient statistic $T(x_i)$ it follows directly from the Fisher-Pearson factorisation that the ratio

$$L(\boldsymbol{\theta}|x_i)/L(\boldsymbol{\theta}|y_i) = k(x_i)/k(y_i)$$

and thus is constant with regard to $\boldsymbol{\theta}$. Consequently, all **data sets in $\mathcal{X}_t$ are also likelihood equivalent**. However, the converse is not true: depending on the sufficient statistics there usually will be many likelihood equivalent data sets that are not part of the same set $\mathcal{X}_t$.

Of particular interest is therefore to find those sufficient statistics that achieve the coarsest partitioning of the sample space and thus may allow the highest data reduction. Specifically, a **minimal sufficient statistic** is a sufficient statistics $T(x_i)$ for which all likelihood equivalent data sets also are equivalent under $T(x_i)$. Therefore, to check whether a sufficient statistic $T(x_i$ is minimally sufficient we verify whether for any two likelihood equivalent data sets x_i and y_i it also follows that $T(x_i) = T(y_i)$. If this holds true then $T(x_i)$ is a minimally sufficient statistic.

An equivalent non-operational definition is that a minimal sufficient statistic $T(x_i)$ is a sufficient statistic that can be computed from any other sufficient statistic $S(x_i)$. This follows from the above directly: assume any sufficient statistic $S(x_i)$, this defines a corresponding set $\mathcal{X}_s$ of likelihood equivalent data sets. By implication any $x_i, y_i \in \mathcal{X}_s$ will ncecessarily also be in $\mathcal{X}_t$, thus whenever $S(x_i) = S(y_i)$ we also have $T(x_i) = T(y_i)$, and therefore $T(x_i)$ is a function of $S(x_i)$.

A trivial but **important example of a minimal sufficient statistic is the likelihood function itself** since by definition it can be computed from any set of sufficient statistics. Thus the likelihood function $L(\boldsymbol{\theta})$ captures all information about $\boldsymbol{\theta}$ that is available in the data. In other words, it provides an *optimal summary* of the observed data with regard to a model. Note that in Bayesian statistics (to be discussed in part 2 of the module) the likelihood function is used as proxy/summary of the data.

For another common example consider the normal model $N(\mu, \sigma^2)$ with parameter vector $\boldsymbol{\theta} = (\mu, \sigma^2)^T$ and log-likelihood

$$l_n(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

One possible set of minimal sufficient statistics for $\boldsymbol{\theta}$ are $\bar{x}$ and $\overline{x^2}$, and with these we can rewrite the log-likelihood function without any reference to the original data x_i as follows

$$l_n(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{n}{2\sigma^2}(\overline{x^2} - 2\bar{x}\mu + \mu^2)$$

An alternative set of minimal sufficient statistics for $\boldsymbol{\theta}$ consists of $s^2 = \overline{x^2} - \bar{x}^2 = \widehat{\sigma^2}_{ML}$ as and $\bar{x} = \hat{\mu}_{ML}$. Note that the dimension of the parameter vector $\boldsymbol{\theta}$ equals the dimension of the minimal sufficient statistic.

Intringuingly **in the normal example the MLEs of the parameters are minimal sufficient statistics**. This is a very useful result that holds true more generally: in the exponential family (which contains the normal distribution as special case) the MLEs of the natural parameters are minimal sufficient statistics.

However, outside the exponential family the MLE is not necessarily a minimal sufficient statistic, and may not even be a sufficient statistic. This is because **a**

(minimal) sufficient statistic of the same dimension as the parameters does not always exist. A classic example is the Cauchy distribution for which the minimal sufficient statistics are the ordered observations, thus the MLE of the parameters do not constitute sufficient statistics. However, the MLE is of course still a function of the minimal sufficient statistic.

In summary, the likelihood function acts as perfect data summariser (i.e. as minimally sufficient statistic), and in exponential families (e.g. Normal distribution) the MLEs of the parameters $\hat{\boldsymbol{\theta}}_{ML}$ are minimimally sufficient.

Finally, while sufficiency is clearly a useful concept for data reduction one needs to keep in mind that this is always in reference to a specific model Therefore, unless one strongly believes in a certain model it is generally a good idea to keep (and not discard!) the original data.

6.4 Summary and concluding remarks on maximum likelihood

6.4.1 Starting point: KL divergence

Finding the model $F_{\boldsymbol{\theta}}$
that best approximates the
underlying true model F_0

$$\rightarrow \text{minimise KL divergence (relative entropy)} \quad D_{\mathrm{KL}}(F_0, F_{\boldsymbol{\theta}})$$

In **large samples** we may approximate
F_0 by the empirical distribution $\hat{F}_0$:

$$\overset{n\to\infty}{\approx} D_{\mathrm{KL}}(\hat{F}_0, F_{\boldsymbol{\theta}})$$

This leads directly to the
method of maximum likelihood!

$$= C - \underbrace{\frac{1}{n}\underbrace{\sum_{i=1}^{n} \log f(x_i|\boldsymbol{\theta})}_{l_n(\boldsymbol{\theta})}}_{\text{cross entropy}}$$

Excursion: Different types of projections: Since the KL divergence is not symmetric there are two ways to minimise the divergence between a fixed F_0 and the optimised $F_{\boldsymbol{\theta}}$, each with different properties:

"forward KL", "approximation KL":	$\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(F_0, F_{\boldsymbol{\theta}})$	"M (Moment) projection" has **zero avoiding** property $f_{\boldsymbol{\theta}}(x) > 0$ whenever $f_0(x) > 0$
"reverse KL", "inference KL":	$\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(F_{\boldsymbol{\theta}}, F_0)$	"I (Information) projection" has **zero forcing** property $f_{\boldsymbol{\theta}}(x) = 0$ whenever $f_0(x) = 0$

6.4.2 Connections between KL divergence, likelihood and expected and observed Fisher information

KL divergence $D_{\mathrm{KL}}(F_0, F_{\boldsymbol{\theta}})$	$\longrightarrow$ local approximation	Expected Fisher Information → property of the model (metric tensor)
$\downarrow$ n large, $F_0 \approx \hat{F}_0$		
log-likelihood function	$\longrightarrow$ local approximation around $\hat{\boldsymbol{\theta}}_{ML}$	Observed Fisher Information → Conditioned an actual observed related to asymptotic variance

(6.1)

It is important to realise that maximum likelihood is only a valid procedure for large sample size n!

6.4.3 Likelihood estimation

Point estimation by maximising the likelihood:

- yields point estimator with many favourable properties.
- most of the time, asymptotic normality allows to assign an error (variance) to the estimate.

Regularity conditions:

To guarantee optimality some regularity conditions need to be met, such as:

- log-likelihood twice differentiable → quadratic approximation near true $\boldsymbol{\theta}$ → normality (log-likelihood of normal is quadratic)
- Fisher information well defined and invertible
- Parameters not on boundary

6.4.4 What happens if n is small?

- Likelihood will *overfit*!

Alternative methods need to be used:

- regularised/penalised likelihood
- Bayesian methods

which are essentially two sides of the same coin.

Classic example of a simple non-ML estimator that is better than the MLE: **Stein's example / Stein paradox** (C. Stein, 1955):

- Problem setting: estimation of the mean in multivariate case
- Maximum likelihood estimation breaks down! → average (=MLE) is worse in terms of MSE than Stein estimator.

6.4.5 Inference with likelihood:

- The likelihood function also allows for inference by the construction of likelihood CIs and corresponding tests.
- Useful for model exploration and model building.

Confidence intervals:

- sets of models that are not statistically distinguishable from the best ML model
- in doubt, choose the simplest model
- better prediction, avoids overfitting

Statistical testing:

- Wald test statistic: equivalent to approximate normal CI
- LRT/GLRT: equivalent to likelihood based CI
- Typically correspond to best available tests

Limits:

GLRT are good for finding a suitable test statistic

- but the χ^2 distribution is only asymptotic, with a better distributions available, e.g., by simulation (parametric bootstrap)
- For comparison of non-nested models it's better to go back to KL divergence $\rightarrow$ AIC information criterion (Akaike, 1972) with a penalised likelihood
- Model selection in small samples and high dimension is challenging $\rightarrow$ current research!

Part II

Bayesian Statistics

Chapter 7

Essentials of Bayesian statistics

7.1 Bayes' theorem

Bayesian statistical learning is linked with the name of Thomas Bayes (1701-1761) who was the first to give Bayes' theorem (1763) on conditional probability.

$$\Pr(A|B) = \Pr(B|A)\frac{\Pr(A)}{\Pr(B)}$$

This theorem relates the two possible conditional probabilities for two events A and B.

7.2 Principle of Bayesian learning

Ingredients:

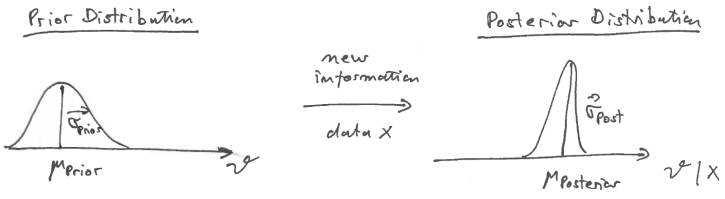
- θ parameter of interest, unknown and fixed.
- prior distribution $\Pr(\theta)$ describing the *uncertainty* (not randomness!) about θ
- data generating process $\Pr(x|\theta)$ (likelihood!)

Question: new information in the form of new observation x arrives - how does the uncertainty about θ change?

Answer: use Bayes' theorem to **update prior distribution to posterior distribution**.

$$\underbrace{\Pr(\theta|x)}_{\text{posterior}} = \frac{\Pr(x|\theta)}{\Pr(x)}\underbrace{\Pr(\theta)}_{\text{prior}}$$

Note that this update procedure can be repeated again and again: we can use the posterior as our new prior and then update it with further data.



For the denominator in Bayes formula we need to compute $\Pr(x)$. This is obtained by

$$\begin{aligned}\Pr(x) &= \mathrm{E}_{F_\theta}\Pr(x|\theta) \\ &= \int \Pr(x|\theta)\Pr(\theta)d\theta \\ &= \int \Pr(x,\theta)d\theta\end{aligned}$$

i.e. by marginalisation of the parameter θ from the joint distribution of θ and x. (For discrete θ replace the integral by a sum).

Depending on the context this quantity is either called *marginal likelihood* (of the underlying model) or *prior predictive distribution* (for the data).

Intringuingly, to conduct a Bayesian statistical analysis typically require integration and/or averaging (e.g. to compute the marginal likelihood), in contrast to maximum likelihood that requires optimisation (to find the maximum likelihood).

7.3 What is exactly is the "Bayesian estimate"?

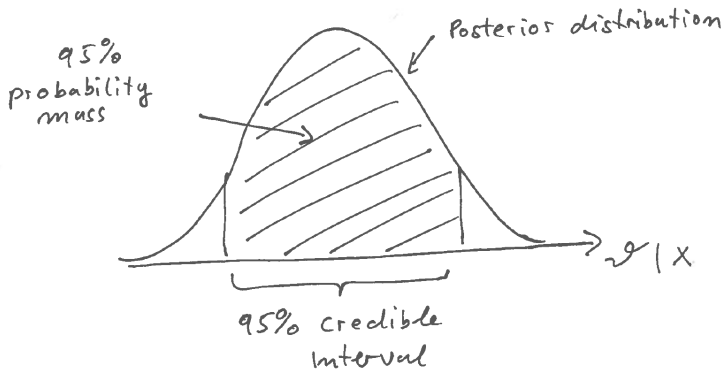
The Bayesian estimate is the full complete posterior distribution!

However, it is useful to summarise aspects of the posterior distribution:

- Posterior mean $\mathrm{E}(\theta|x)$
- Posterior variance $\mathrm{Var}(\theta|x)$
- Posterior mode etc.

In particular the mean of the posterior distribution is often taken as a *Bayesian point estimate*.

The posterior distribution also allows to define **credible regions** or **credible intervals**. These are the **Bayesian equivalent to confidence intervals** and are constructed by finding the areas of highest probability mass (say 95%) in the posterior distribution.



Bayesian credible intervals (unlike their frequentist confidence counterparts) are thus very easy to interpret - they simply correspond to the area in the parameter space in which the we can find the parameter with a given specified probability. In contrast, in frequentist statistics it does not make sense to assign a probability to a parameter value!

Note that there are typically many credible intervals with the given specified coverage α (say 95%). Therefore, we may need further criteria to constuct these intervals.

In the univariate case a **two-sided equal-tail credible interval** is obtained by finding the corresponding lower $1 - \alpha/2$ and upper $\alpha/2$ quantiles.

A **highest posterior density (HDP)** interval of coverage α is found by identifying the shortest interval (i.e. with smallest support) for the given α probability mass. Any point within such an interval has higher density resp. probability than outside the credible interval. When the posterior has multiple peaks this means the the HPD interval may be disjoint.

7.4 Computer implementation of Bayesian learning

As we have seen Bayesian learning is *conceptually straightforward*:

1) Specify prior uncertainty $\Pr(\theta)$ about the parameters of interest θ.
2) Specify the data generating process for a specified parameter: $\Pr(x|\theta)$.
3) Apply Bayes' theorem to update prior uncertainty in the light of the new data.

In practise, however, computing the posterior distribution can be *computationally very demanding*, especially for complex models.

For this reason specialised software packages have been developed for computational Bayesian modelling, for example:

- Stan probabilistic programming language (can be used with R and Python) — https://mc-stan.org/
- PyMC3 and PyMC4 (Python) — https://docs.pymc.io/
- TensorFlow Probability / Edward2 (Python) — https://www.tensorflow.org/probability/

- BUGS for Bayesian analysis — https://www.mrc-bsu.cam.ac.uk/software/bugs/ .

7.5 Bayesian interpretation of probability

7.5.1 What makes you "Bayesian"?

If you use Bayes' theorem are you therefore automatically a Bayesian? No!!

Bayes' theorem is a mathematical fact from probability theory. Hence, Bayes' theorem is valid for everyone, whichever form for statistical learning your are subscribing (such as frequentist ideas, likelihood methods, entropy learning, Bayesian learning).

As we discuss now the key difference between Bayesian and frequentist statistical learning lies in the differences in *interpretation of probability*, not in the mathematical formalism for probability (which includes Bayes' theorem).

7.5.2 Mathematics of probability

The mathematics of probability in its modern foundation was developed by Kolmogorov (1933). Essentially, this establishes probability in terms of set theory/measure theory. This theory provides a coherent mathematical framework to work with probabilities (see module "Foundations of Modern Probability").

However, Kolmogorov's theory does *not* provide an interpretation of probability!

$\rightarrow$ The Kolmogorov framework is the basis for both the frequentist and the Bayesian interpretation of probability.

7.5.3 Interpretations of probability

Essentially, there are two major commonly used interpretation of probability in statistics - the **frequentist interpretation** and the **Bayesian interpretation**.

7.5.3.1 A: Frequentist interpretation

probability = frequency (of an event in a long-running series of identically repeated experiments)

This is the *ontological view* of probability (i.e. probability "exists" and is identical to something that can be observed.).

It is also a very restrictive view of probability. For example, frequentist probability cannot be used to describe events that occur only a single time. Frequentist probability thus can only be applied asymptotically, for large samples!

7.5.3.2 B: Bayesian probability

"Probability does not exist" (famous quote by Bruno de Finetti, a Bayesian statistician)

What does this mean?

Probability is a **description of the state of knowledge** and of **uncertainty**.

Probability is thus an *epistemological quantity* that is assigned and that changes rather than something that is an inherent property of an object.

Note that this does not require any repeated experiments. The Bayesian interpretation of probability is valid regardless of sample size or the number or repetitions of an experiment.

Hence, the key difference between frequentist and Bayesian approaches is not the use of Bayes' theorem. Rather it is whether you consider probability as ontological (frequentist) or epistemological entity (Bayesian).

7.6 Historical developments

- Thomas Bayes (1701-1761) the father of Bayesian statistics
  Only after his death his paper on Bayes' theorem was published (1763).
- Laplace (from 1800) was actually the first to use Bayes' theorem for statistical calculations. This activivity was then called "inverse probability".
- Between 1900 and 1940 classical mathematical statistics was developed and the field was heavily influenced and dominated by R.A. Fisher (who invented likelihood theory and ANOVA, among other things - he also was working in population genetics). Fisher himself was very much opposed to Bayesian theory.
- 1931 de Finetti publishes his "representation theorem". This shows that the joint distribution of a sequence of exchangeable events (i.e. where the ordering can be permuted) can be represented by a mixture distribution that can be constructed via Bayes' theorem. (Note that exchangeability is a weaker condition than i.i.d.) This theorem is often used as a justification Bayesian statistics (along with the socalled Dutch book argument, also by de Finetti).
- 1933 publication of Kolmogorov's book on probability theory.
- 1946 Cox theorem (R. T. Cox): the aim to generalise classical logic (from TRUE/FALSE to continuous measures of uncertainty) inevitably leads to probability theory and Bayesian learning! This justification of Bayesian statistics was later popularised by E.T. Jaynes in various books (1959, 2003).
- 1955 Stein Paradox - Charles Stein publishes paper on the Stein estimator - an estimator of the mean that dominates ML estimator. His estimator is always better in terms of MSE than the ML estimator, and this was very puzzling at that time!

From 1970 onwards Bayesian learning has become more pervasive!

- Computers allow to do the complex computations needed in Bayesian statistics
- Metropolis-Hastings algorithm published

- A lot of work on interpreting Stein estimators as empirical Bayes estimators (Efron and Morris 1975) and on development of regularised estimation techniques such as penalised likelihood in regression (e.g. ridge regression)
- regularisation originally was only meant to make singular systems/matrices invertible - but then it turned out regularisation has a simple Bayesian interpretation!
- work on reference priors (Bernado 1979)
- penalised likelihood via KL divergence for model selection (Akaike 1973)

Another boost was in the 1990/2000s when in science (e.g. genomics) many complex and high-dimensional data set were becoming widely available. Classical statistical methods cannot be used in this setting (overfitting!) so many new methods were developed for high-dimensional data analysis, many with direct link to Bayesian statistics:

- 1996 lasso regression (Tibshirani)
- Machine learning etc (many Bayesians in this field!)

In short, Bayesian statistics has many favourable properties:

- applicable to small samples (and even to single events!)
- automatically regularises (via the prior) which is important for complex models and when there is the problem of overfitting.
- it provides a coherent generalisation of classical TRUE/FALSE logic
- it is conceptually very simple (but computationally more involved)
- asympotically (large n) it is consistent and converges to the true model (like ML!).

7.7 Connection with entropy learning

7.7.1 Zero forcing property

It is easy to see that if in Bayes rule the prior probability for an event is set to 0, then the posterior probability for that event will remain at 0, regardless of the data! This **zero-forcing property** of the Bayes update rule has been called **Cromwell's rule** by D. Lindley. Therefore, assigning prior probability 0 to an event should be avoided.

Note that this implies that assigning prior probability 1 to an event should be avoided, too, since this means assigning 0 to all other alternative events.

7.7.2 Connection with entropy learning

The *Bayesian update rule* is a very general form of learning when the *new information arrives in the form of data*.

But actually there is an even a more general principle: the **principle of minimal information update** (e.g. Jaynes 1959, 2003) or **principle of minimum information discrimination (MDI) (Kullback 1959)**:

- **Change your beliefs only as much as necessary to be coherent with new evidence!**

This is also called **entropy learning** since the KL divergence $(F_{\theta|\text{new information}}; F_\theta)$ is employed to measure the divergence from the updated distribution to the distribution prior to the arrival of the information.

Note that this update is based on an I-projection (see Part I, Likelihood), which also does have the zero forcing property (hinting that Bayes rule is a special case).

Thus, when new information arrives then the uncertainty about the parameter is only minimally adjusted, just as much as needed to account for the new information ("inertia of beliefs").

There are three main special cases that follow from the entropy learning rule:

1) if information arrives in form of data $\rightarrow$ update by T. Bayes' theorem (1763)
2) if information is in the form of another distribution $\rightarrow$ update using R. Jeffrey's rule (1965)
3) if the information is in form of constraints $\rightarrow$ Kullback's principle of minimum MDI (1959), E. T. Jaynes MaxEnt principle (1957)

Since 1) is by far the most common situation it is clear why it is important to study Bayesian learning!

This shows (again) how fundamentally important KL divergence is in statistics - it not only leads to likelihood inference but also to Bayesian learning, as well as to other forms of information updating! Furthermore, relative entropy is useful to choose priors (e.g. reference priors) and in experimental design.

Chapter 8

Beta-Binomial model for estimating a proportion

In this chapter we discuss how to estimate a portion in the Bayesian framework.

8.1 Binomial likelihood

In order to apply Bayes' theorem we first need to find a suitable likelihood based on modeling the data generating process. Here we follow the Binomial model as used previously in Part I:

Repeated Bernouli experiment (Binomial model):

$x \in \{0, 1\}$ (e.g. "tails" vs. "heads")

proability mass function (pmf): $\Pr(x = 1) = p$, $\Pr(x = 0) = 1 - p$
Mean: $\mathrm{E}(x) = p$
Variance $\mathrm{Var}(x) = p(1 - p)$

$Binomial(n, p)$ (sum of n Bernoulli experiments)
$x \in \{0, 1, \ldots, n\}$
Mean: $\mathrm{E}(x) = np$
Variance: $\mathrm{Var}(x) = np(1 - p)$

Standardised Binomial (average of n Bernoulli experiments):
$\frac{x}{n} \in \{0, \frac{1}{n}, \ldots, 1\}$
Mean: $\mathrm{E}(\frac{x}{n}) = p$
Variance: $\mathrm{Var}(\frac{x}{n}) = \frac{p(1-p)}{n}$

From part I (likelihood theory) we know that the *maximum likelihood estimate* of the proportion is the frequency $\hat{p}_{ML} = \frac{x}{n}$ given x (number of "heads") is observed in n repeats.

8.2 Excursion: Properties of the Beta distribution

The density of the Beta distribution $Beta(\alpha, \beta)$ for $x \in [0, 1]$ and $\alpha > 0$ and $\beta > 0$ is

$$f(x|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

The mean is $\mathrm{E}(x) = \mu = \frac{\alpha}{\alpha+\beta}$ and the variance $\mathrm{Var}(x) = \frac{\mu(1-\mu)}{\alpha+\beta+1}$.

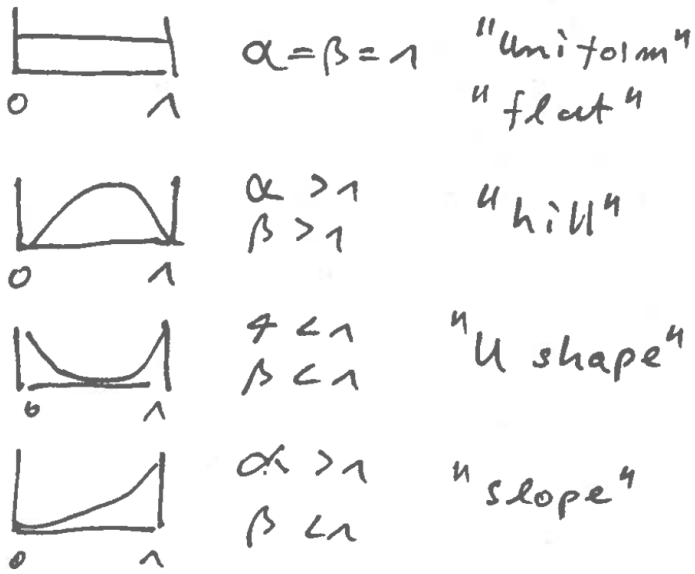
The density depends on the Beta function $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ which in turn is defined via Euler's Gamma function

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$$

Note that $\Gamma(x) = (x-1)!$ for any positive integer x

A useful reparameterisation of the Beta distribution is in terms of the parameters $\mu \in [0, 1]$ and $m > 0$, yielding the original parameters via $\alpha = \mu m$ and $\beta = (1-\mu)m$. Conversely, $m = \alpha + \beta$ and $\mu = \frac{\alpha}{\alpha+\beta}$.

The Beta distribution is very flexible and can assume a number of different shapes, depending on the value of α and β:



8.3 Beta prior distribution

In Bayesian learning we need to make explicit our uncertainty about p.

p has support $[0, 1] \rightarrow$ we use the **Beta distribution** $Beta(\alpha, \beta)$ **as prior for** p with parameters $\alpha \geq 0$ and $\beta \geq 0$:

$$p \sim Beta(\alpha, \beta)$$

Note this does not actually mean that p is random! It only means that we model the uncertainty about p using a Beta random variable!

The flexibility of the Beta distribution allows to accomodate a large variety of possible scenarious for our prior knowledge.

The prior mean is

$$\mathrm{E}(p) = \frac{\alpha}{m} = \mu_{\text{prior}}$$

and the prior variance

$$\mathrm{Var}(p) = \frac{\mu_{\text{prior}}(1-\mu_{\text{prior}})}{m+1}$$

where $m = \alpha + \beta$.

Note the similarity to the moments of the standardised Binomial above!

8.4 Computing the posterior distribution

Bayes' theorem for continous random variables to compute posterior density:

$$f(p|x) = \frac{f(x|p)f(p)}{\int_{p'} f(x|p')f(p')dp'}$$

We use in our analysis the Beta-Binomial model:

a) **Beta prior:**

$$p \sim Beta(\alpha, \beta)$$
$$f(p) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

b) **Binomial likelihood:**

$$x|p \sim Binom(n, p)$$
$$f(x|p) = \binom{n}{x} p^x (1-p)^{(n-x)}$$

Applying Bayes' theorem results in

c) **Beta posterior distribution**

$$p|x \sim Beta(\alpha + x, \beta + n - x)$$
$$f(p|x) = \frac{1}{B(\alpha + x, \beta + n - x)} p^{\alpha+x-1}(1-p)^{\beta+n-x-1}$$

(for a proof see Worksheet 5!)

The posterior can be summarised by its first two moments (mean and variance):

Posterior mean:

$$\mu_{\text{posterior}} = \mathrm{E}(p|x) = \frac{x+\alpha}{n+m}$$

Posterior variance:

$$\sigma^2_{\text{posterior}} = \mathrm{Var}(p|x) = \frac{\mu_{\text{posterior}}(1-\mu_{\text{posterior}})}{n+m+1}$$

Chapter 9

Properties of Bayesian learning

The Beta-Binomial models allows to observe a number of intriguing features and properties of Bayesian learning. Many of these extend also to other models as we will see later.

9.1 Prior acting as pseudo-data

In the expression for the posterior mean and variance you can see that $m = \alpha + \beta$ behaves like an implicit sample size connected with prior information!

Specifically, α and β act as **pseudo-counts** that influence both the posterior mean and the posterior variance, exactly in the same way as coventional data.

For example, larger m (and thus α and β) the smaller is the posterior variance, with variance decreasing proportional to the inverse of m. If the prior is highly concentrated, i.e. if it has low variance and large precision (=inverse variance) then the implicit data size m is large. Conversely, if the prior has a large variance, then the prior is vague and the implicit data size m is small.

Hence, a prior has the same effect as if one would add data – but without actually adding data! This is precisely this why a prior acts as a regulariser and prevents overfitting, because it increases effective sample size.

Another interpretation is that any prior summarises data that may have been available previously as observations.

9.2 Linear shrinkage of mean

The posterior mean $\mu_{\text{posterior}}$ is a linearly adjusted $\hat{\mu}_{ML}$. This becomes evident by writing $\mu_{\text{posterior}}$ as

$$\mu_{\text{posterior}} = \lambda \mu_{\text{prior}} + (1 - \lambda)\hat{\mu}_{ML}$$

with weight $\lambda \in [0, 1]$

$$\lambda = \frac{m}{m+n}.$$

The **posterior mean is a convex combination (i.e. the weighted average) of the ML estimate and the prior mean**. The factor λ is called the **shrinkage intensity** — note that it is the ratio of the "prior sample size" (m) and the "effective overall sample size" ($m + n$).

1. This is called *shrinkage*, because the ML estimator is "shrunk" towards the prior mean (which is often called the "target", and sometimes the target is zero, and then the terminology "shrinking" makes most sense).

2. If the shrinkage intensity is zero ($\lambda = 0$) then the ML point estimator is recovered. This implies $\alpha = 0$ and $\beta = 0$, or $n \to \infty$.

   Note that using maximum likelihood to estimate of the proportion p (for moderate or small n) is the same as Bayesian estimation using the Beta-Binomial model with prior $\alpha = 0$ and $\beta = 0$. This prior is extremely "u-shaped" and the implicit prior for the ML estimation. (Would you would use such a prior intentionally?)

3. If the shrinkage intensity is large ($\lambda \to 1$) then the posterior mean corresponds to the prior. This happens if $n = 0$ or if m is very large (implying that the prior is sharply concentrated around the prior mean).

4. Since the ML estimate (=frequency) is unbiased the Bayesian point estimate is biased (for finite n)! And the bias is in fact the prior mean! So Bayesian statistics produces by default biased estimators (but asymptotically they will be unbiased like in ML).

5. That the posterior mean is a linear combination of the ML estimate and the prior mean is not a coincidence. In fact, this is true for all distributions in the exponential family (see e.g. Diaconis and Ylvisaker, 1979). Furthermore, it is possible (and indeed quite useful for computational reasons!) to formulate Bayes theory completely in terms of linear shrinkage (e.g. Hartigan 1969). The resulting theory is called "Bayes linear statistics" (Goldstein and Wooff, 2007).

9.3 Conjugacy of prior and posterior distribution

In the Beta-Binomial model for estimating the proportion p the choice of the **Beta distribution as prior distribution** along with the Binomial likelihood resulted in having the **Beta distribution as posterior distribution** as well.

If the prior and posterior belong to the same distributional family the prior is called a **conjugate prior**. This will be the case if the prior has the same functional form as the likelihood.

In the Beta-Binomial the likelihood is based on the Binomial distribution and has the following form (only terms depending on the parameter p are shown):

$$p^x(1-p)^{n-x}$$

The form of the Beta prior is (again, only showing terms depending on p):

$$p^{\alpha-1}(1-p)^{\beta-1}$$

Since the posterior is proportional to the product of prior and likelihood the posterior will have exactly the same form as the prior:

$$p^{\alpha+x-1}(1-p)^{\beta+n-x-1}$$

Choosing the prior distribution from a family conjugate to the likelihood greatly simplifies Bayesian analysis since the Bayes formula can then be written in form of an update formula for the parameters of the Beta distribution:

$$\alpha \rightarrow \alpha + x$$

$$\beta \rightarrow \beta + n - x$$

Thus, conjugate prior distributions are very convenient choices. However, in their application it must be ensured that the prior distribution is flexible enough to encapsulate all prior information that may be available. In cases where this is not the case alternative priors should be used (and most likely this will then require to compute the posterior distribution numerically rather than analytically).

9.4 Large sample asymptotics

9.4.1 Large sample limits of mean and variance

If n is large and $n >> \alpha, \beta$ the posterior mean and variance become asympotically

$$\mu_{\text{posterior}} \stackrel{a}{=} \frac{x}{n} = \hat{\mu}_{ML}$$

and

$$\sigma^2_{\text{posterior}} \stackrel{a}{=} \frac{\hat{\mu}_{ML}(1-\hat{\mu}_{ML})}{n}$$

Thus, if sample size is large the Bayes' estimator turns into the ML estimator! Specifically, the posterior mean becomes the ML point estimate, and the posterior variance is equal to the asymptotic variance computed via the observed Fisher information!

Thus, for large n the data dominate and any details about the prior (such as values of α and β become irrelevant!

9.4.2 Asymptotic Normality of the Posterior distribution

Also known as **Bayesian Central Limit Theorem (CLT)**.

Under some regularity conditions (such as regular likelihood and positive prior probability for all parameter values, finite number of parameters, etc.) for

large sample size the Bayesian posterior distribution converges to a Normal distribution centered around the MLE and with the variance of the MLE:

$$\text{for large } n\text{: } \Pr(\boldsymbol{\theta}|\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) \rightarrow N(\hat{\boldsymbol{\theta}}_{ML}, \text{Var}(\hat{\boldsymbol{\theta}}_{ML}))$$

So not only are the posterior mean and variance converging to the MLE and the variance of the MLE for large sample size, but also the posterior distribution itself converges to the sampling distribution!

This holds generally in many regular cases, not just in our example of the Beta-Bernoulli model.

The Bayesian CLT is generally known as the **Bernstein-van Mises theorem** (who discovered it at around 1920-30), but special cases were already known as by Laplace.

In the Worksheet 5 the asymptotic convergence of the posterior distribution to a normal distribution is demonstrated grapically.

9.5 Posterior variance for finite n

In the previous chapter we have the derived a Bayesian point estimate for the proportion p as the posterior mean

$$\text{E}(p|x) = \frac{x + \alpha}{n + m} = \hat{p}_{\text{Bayes}}$$

with posterior variance

$$\text{Var}(p|x) = \frac{\hat{p}_{\text{Bayes}}(1 - \hat{p}_{\text{Bayes}})}{n + m + 1}$$

Asymptotically, we have seen that for large n the posterior becomes the ML estimator, and the posterior variance becomes the asymptotic variance of the MLE. Thus, the Bayesian estimate will be indistinguishable from the MLE for large n and shares its favourable properties.

In addition, for finite sample size the posterior variance will tyically be *smaller* than both the asymptotic posterior variance (for large n) and the prior variance, showing that combining the information in the prior and in the data leads to a more effecient estimate.

Chapter 10

Normal-Normal and Inverse-Gamma-Normal models for estimating the mean and the variance

10.1 Normal-Normal model to estimate mean

10.1.1 Normal likelihood

For the **likelihood** we assume as data-generating model the normal distribution with known fixed variance σ^2

$$x|\mu \sim N(\mu, \sigma^2)$$

This yields as the MLE $\hat{\mu}_{ML} = \bar{x}$.

10.1.2 Normal prior distribution

To model the uncertainty about μ we use the normal distribution $N(\mu, \sigma^2/k)$ parameterised by the two parameters μ and k (remember σ^2 is fixed).

With $\mu = \mu_0$ and $k = m$ we get the **normal prior**

$$\mu \sim N(\mu_0, \sigma^2/m)$$

with prior mean $\mathrm{E}(\mu) = \mu_0$ and prior variance $\mathrm{Var}(\mu) = \frac{\sigma^2}{m}$ where m is the implied sample size from the prior. Note that m does not need to be an integer value!

10.1.3 Normal posterior distribution

The **posterior distribution** after observing n samples $x_1, \dots x_n$ is normal with $\mu = \mu_1$ and $k = m + n$

$$\mu | x_1, \dots x_n \sim N(\mu_1, \sigma^2/(m+n))$$

with posterior mean

$$\mathrm{E}(\mu | x_1, \dots x_n) = \mu_1 = \frac{m\mu_0 + n\bar{x}}{n+m} = \lambda\mu_0 + (1-\lambda)\hat{\mu}_{ML}$$

with $\lambda = \frac{m}{n+m}$. Note the linear shrinkage of $\hat{\mu}_{ML}$ towards μ_0!

The corresponding posterior variance is

$$\mathrm{Var}(\mu | x_1, \dots x_n) = \frac{\sigma^2}{n+m}$$

Thus, the **normal distribution is the conjugate distribution to the mean parameter in the normal likelihood**.

10.1.4 Large sample asymptotics and Stein paradox

For n large and $n >> m$ we get

$$\mathrm{E}(\mu | x_1, \dots x_n) \overset{a}{=} \hat{\mu}_{ML}$$

$$\mathrm{Var}(\mu | x_1, \dots x_n) \overset{a}{=} \frac{\sigma^2}{n}$$

i.e. the MLE and its asymptotic variance!

Note that the posterior variance $\frac{\sigma^2}{n+m}$ is smaller than the asymptotic variance $\frac{\sigma^2}{n}$ and the prior variance $\frac{\sigma^2}{m}$.

When studying the frequentist properties of the posterior mean μ_1 it turns out that by an appropriate choice of m (or λ) it is possible to construct an estimator that will outperform the MLE for finite n in terms of MSE (with the reduced variance compensating for the increase in bias)! Charles Stein was one of the first to present such an estimator (see next chapter), and by many of his contemporaries it was considered very puzzling to have any estimator outperform the MLE, hence this effect is called **Stein paradox**.

10.2 Inverse-Gamma-Normal model to estimate variance

10.2.1 Inverse Gamma distribution

Next, we study a common Bayesian model for estimating the variance parameter of the normal distribution. For this we use the inverse Gamma distribution:

$$x \sim IG(\alpha, \beta)$$

This distribution is closely linked with the Gamma distribution — the inverse of x is Gamma-distributed with inverted scale parameter:

$$\frac{1}{x} \sim \text{Gamma}(\alpha, \beta^{-1})$$

For use as prior and posterior we employ a different parameterisation with $k = 2(\alpha - 1)$ and $v = \beta/(\alpha - 1)$:

$$x \sim IG(1 + \frac{k}{2}, \frac{k}{2}v)$$

The first two moments of the IG distribution are

$$\text{E}(x) = \frac{\beta}{\alpha - 1} = v$$

and

$$\text{Var}(x) = \frac{\beta^2}{(\alpha - 1)^2(\alpha - 2)} = \frac{2v^2}{k - 2}$$

10.2.2 Normal likelihoood

As data likelihood / generating model we use normal distribution $N(\mu, \sigma^2)$ with given fixed mean μ.

This yields as MLE $\widehat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$

10.2.3 Inverse Gamma prior distribution

For the prior distribution we use the inverse Gamma distribution with with $k = m$ and $v = \sigma_0^2$

$$\sigma^2 \sim IG(k = m, v = \sigma_0^2)$$

The corresponding prior mean is

$$\text{E}(\sigma^2) = \sigma_0^2$$

and the prior variance is

$$\text{Var}(\sigma^2) = \frac{2\sigma_0^4}{m - 2}$$

(note that $m > 2$)

10.2.4 Inverse Gamma posterior distribution

As the inverse Gammma distribution is conjugate to the normal likelihood the posterior distribution is inverse Gamma as well:

$$\sigma^2 | x_1 \ldots, x_n \sim IG(k = m + n, v = \sigma_1^2)$$

with $\sigma_1^2 = \frac{\sigma_0^2 m + n\widehat{\sigma}^2_{ML}}{m+n}$.

The posterior mean is

$$\mathrm{E}(\sigma^2|x_1\ldots,x_n) = \sigma_1^2$$

and the posterior variance

$$\mathrm{Var}(\sigma^2|x_1\ldots,x_n) = \frac{2\sigma_1^4}{m+n-2}$$

The update formula for the posterior mean of the variance follows the usual linear shrinkage rule:

$$\sigma_1^2 = \lambda\sigma_0^2 + (1-\lambda)\widehat{\sigma}_{ML}^2$$

with $\lambda = \frac{m}{m+n}$.

10.2.5 Large sample asymptotics

For n large and $n >> m$ we get

$$\mathrm{E}(\sigma^2|x_1,\ldots x_n) \stackrel{a}{=} \widehat{\sigma}_{ML}^2$$

$$\mathrm{Var}(\sigma^2|x_1,\ldots x_n) \stackrel{a}{=} \frac{2\sigma^4}{n}$$

which is indeed the MLE of σ^2 and its asymptotic variance!

10.2.6 Estimating precision

Instead of estimating the variance it is actually a bit simpler to estimate the precision (i.e. the inverse variance). For this one would then use a Gamma prior and a normal likelihood, resulting in a Gamma posterior.

10.2.7 Joint estimation of mean and variance

It is possible to combine the Normal-Normal for the mean and the Inverse-Gamma-Normal model into a joint model for the mean and variance.

This implies having a joint prior and a joint posterior for μ and σ^2.

Details are not shown here but the resulting joint point estimators are identical to the above individual estimators.

Chapter 11

Shrinkage estimation using empirical risk minimisation

11.1 Linear shrinkage

In the examples for Bayesian estimation we have seen so far the posterior mean of the parameter of interest was obtained by linear shrinkage

$$\hat{\theta}_{\text{shrink}} = \mathrm{E}(\theta|x_1, \ldots, x_n) = \lambda\theta_0 + (1-\lambda)\hat{\theta}_{\text{ML}}$$

of the MLE $\hat{\theta}_{\text{ML}}$ towards the prior mean θ_0, with shrinkage intensity $\lambda = \frac{m}{m+n}$ determined by the pseudo-sample size m (which in turn is linked the precision of the prior) and the sample size n.

The resulting point estimate $\hat{\theta}_{\text{shrink}}$ is called *shrinkage estimate* and is a convex combination of θ_0 and $\hat{\theta}_{\text{ML}}$. The prior mean θ_0 is also called the "target".

In a Bayesian estimation the parameter m and hence λ is given a priori, but it turns out that it is possible and useful to find an optimal value for λ by minimising the mean squared error of the estimator $\hat{\theta}_{\text{shrink}}$.

In particular, by construction, the target θ_0 has zero variance but substantial bias, whereas the MLE $\hat{\theta}_{\text{ML}}$ will have low or zero bias but a non-vanishing variance. By combinining these two estimators with opposite properties the aim is to achieve a *bias-variance tradeoff* so that the resulting estimator $\hat{\theta}_{\text{shrink}}$ has lower MSE than either θ_0 and $\hat{\theta}_{\text{ML}}$.

Specifically, the aim is to find

$$\lambda^\star = \arg\min_\lambda \mathrm{E}\left((\theta - \hat{\theta}_{\text{shrink}})^2\right)$$

It turns out that this can be minimised without knowing the actual true value of θ and the result for an unbiased $\hat{\theta}_{\text{ML}}$ is

$$\lambda^\star = \frac{\mathrm{Var}(\hat{\theta}_{\text{ML}})}{\mathrm{E}((\hat{\theta}_{\text{ML}} - \theta_0)^2)}$$

Hence, the shrinkage intensity will be small if the variance of the MLE is small and/or if the target and the MLE differ substantially. On the other hand, if the variance of the MLE is large and/or the target is close to the MLE the shrinkage intensity will be large.

11.2 James-Stein estimator

We can now use empirical risk minimisation to estimate the shrinkage parameter the Normal-Normal model.

In 1955 James and Stein propose the following estimate for the multivariate mean $\boldsymbol{\mu}$ of using a single sample $\boldsymbol{x}$ drawn from the multivariate normal $N_d(\boldsymbol{\mu}, \boldsymbol{I})$:

$$\hat{\boldsymbol{\mu}}_{JS} = (1 - \frac{d-2}{||\boldsymbol{x}||^2})\boldsymbol{x}$$

Here, we recognise $\hat{\boldsymbol{\mu}}_{ML} = \boldsymbol{x}$, $\boldsymbol{\mu}_0 = 0$ and shrinkage intensity $\lambda^\star = \frac{d-2}{||\boldsymbol{x}||^2}$.

Efron and Morris (1972) and Lindley and Smith (1972) generalised this shrinkage estimator to the case of multiple observations $\boldsymbol{x}_1, \ldots \boldsymbol{x}_n$ and target $\boldsymbol{\mu}_0$, yielding an empirical Bayes estimate of μ based on the Normal-Normal model.

Chapter 12

Bayesian model comparison using Bayes factors and the BIC

12.1 The Bayes factor

We would like to compare two models M_1 and M_2. Before seeing data D we can check their **Prior odds** (= ratio of prior probabilities of the models M_1 and M_2):

$$\frac{\Pr(M_1)}{\Pr(M_2)}$$

After seeing data D we arrive at the **Posterior odds** (= ratio of posterior probabilities):

$$\frac{\Pr(M_1|D)}{\Pr(M_2|D)}$$

Using Bayes Theorem $\Pr(M_i|D) = \frac{\Pr(D|M_i)\Pr(M_i)}{\Pr(D)}$ we can rewrite the posterior odds as

$$\underbrace{\frac{\Pr(M_1|D)}{\Pr(M_2|D)}}_{\text{posterior odds}} = \underbrace{\frac{\Pr(D|M_1)}{\Pr(D|M_2)}}_{\text{Bayes factor } B_{12}} \underbrace{\frac{\Pr(M_1)}{\Pr(M_2)}}_{\text{prior odds}}$$

The **Bayes factor** is the multiplicative factor that updates the prior odds to the posterior odds, and is the ratio of the (marginal) likelihoods of the two models:

$$B_{12} = \frac{\Pr(D|M_1)}{\Pr(D|M_2)}$$

The **log-Bayes factor** $\log B_{12}$ is also called the **weight of evidence** for M_1 over M_2. Therefore, we see that

log-posterior odds = weight of evidence + log-prior odds

12.1.1 Connection with relative entropy

The *expected* weight of evidence, with expectation taken with regard to one of the two models, is in fact the **KL divergence** between the two models (plus a minus sign depending on direction):

$$\mathrm{E}_{M_1}(\log B_{12}) = KL(M_1||M_2)$$

$$\mathrm{E}_{M_2}(\log B_{12}) = -\mathrm{E}_{M_2}(\log B_{21}) = -KL(M_2||M_1)$$

12.1.2 Interpretation of and scale for Bayes factor

Following Harold Jeffreys (1961) one may interpret the strength of the Bayes factor as follows:

B_{12}	$\log B_{12}$	evidence in favour of M_1 versus M_2
> 100	> 4.6	decisive
10 to 100	2.3 to 4.6	strong
3.2 to 10	1.16 to 2.3	substantial
1 to 3.2	0 to 1.16	not worth more than a bare mention

More recently, Kass and Raftery (1995) proposed to use the following slightly modified scale:

B_{12}	$\log B_{12}$	evidence in favour of M_1 versus M_2
> 150	> 5	very strong
20 to 150	3 to 5	strong
3 to 20	1 to 3	positive
1 to 3	0 to 1	not worth more than a bare mention

12.1.3 Computing $\Pr(D|M)$ for simple and composite models

In the Bayes factor we need to compute $\Pr(D|M)$, and it turns out that this is different for simple and composite models.

A model is called "simple" if it directly corresponds to a specific distribution, say, a Normal with fixed mean and variance, or a Binomial distribution with a set probability for the two classes. Thus, a simple model is a point in the model space described by the parameters of a distribution family (e.g. μ and σ^2 for the normal family $N(\mu, \sigma^2)$. For a simple model M the probability $\Pr(D|M)$ is the likelihood of M.

On the other hand, a model is "composite" if it is composed of simple models. This can be a finite set, or it can be comprised of infinite number of models. For

example, a Normal with a given mean but unspecified variance, or a Binomial model with unspecified parameter p, is a composite model.

If M is a composite model, with the underlying simple models indexed by a parameter θ, the probability of the data given the model is obtained by marginalisation over θ:

$$\begin{aligned}\Pr(D|M) &= \int \Pr(D|\theta, M)\Pr(\theta|M)d\theta \\ &= \int \Pr(D, \theta|M)d\theta\end{aligned}$$

i.e. we *integrate* over all parameter values θ. The resulting probability is called the *marginal likelihood* of the model M. Note the marginal likelihood appears also in the denominator of Bayes formula! The marginal distribution for D is also called the prior predictive distribution given M.

If the distribution over θ is strongly concentrated around a specific value then the composite model degenerates to a simple point model.

A worked example (in the form of the Beta-Binomial distribution) is discussed in more detail in the Worksheet 6, Question 3.

12.1.4 Bayes factor versus likelihood ratio

If both M_1 and M_2 are simple models then the **Bayes factor is identical to the likelihood ratio** of the two models.

However, if one of the two models is composite then the Bayes factor and the generalised likelihood ratio differ: In the Bayes factor the representative of a composite model is the **model average** of the simple models indexed by θ, with weights taken from the prior distribution over the simple models contained in M. In contrast, in contrast in the generalised likelihood ratio statistic the representative of a composite model is chosen by *maximisation*!

Thus, **for composite models**, the **Bayes factor does *not* equal the corresponding generalised likelihood ratio statistic.** As we will see next when studying the BIC approximation, the key difference is that the Bayes factor takes into account the dimension of the composite models.

12.2 Approximate computation of the marginal likelihood and of the log-Bayes factor

The marginal likelihood and the Bayes factor can be difficult to compute in practise. Therefore, a number of approximations for Bayesian modeling and model selection have been developed The most important is the so-called BIC approximation.

12.2.1 Schwarz (1978) approximation of log-marginal likelihood

The logarithm of the marginal likelihood of a model can be approximated using the so-called BIC approximation (Schwarz 1978) as follow:

$$\log \Pr(D|M) \approx l_n^M(\hat{\boldsymbol{\theta}}_{ML}^M) - \frac{1}{2} d_M \log n$$

where d_M is the dimension of the model M (number of parameters in $\boldsymbol{\theta}$ belonging to M) and n is the sample size and $\hat{\boldsymbol{\theta}}_{ML}^M$ is the MLE. For a simple model $d_M = 0$ so then there is no approximation as in this case the marginal likelihood equals the likelihood.

The above formula can be obtained by quadratic approximation of the likelihood **assuming large** n and that the prior is uniform around the MLE.

Note that the approximation is the maximum log-likelihood minus a penalty that depends on the model complexity (as measured by dimension d), thus this is an example of penalised ML! Also note that the distribution over the parameter θ is not required in the approximation.

12.2.2 Bayesian information criterion (BIC)

The BIC (Bayesian information criterion) of the model M is the approximated log-marginal likelihood times the factor -2:

$$BIC(M) = -2 l_n^M(\hat{\boldsymbol{\theta}}_{ML}^M) + d_M \log n$$

Thus, when comparing models one aimes to maximise the marginal likelihood or, as approximation, minimise the BIC.

The reason for the factor "-2" is simply to have a quantity that is on the same scale as the Wilks log likelihood ratio. Some people / software packages also use the factor "2".

12.2.3 Approximating the weight of evidence (log-Bayes factor) with BIC

Using BIC (twice) the log-Bayes factor can be approximated as

$$\begin{aligned} 2 \log B_{12} &\approx -BIC(M_1) + BIC(M_2) \\ &= 2 \left(l_n^{M_1}(\hat{\boldsymbol{\theta}}_{ML}^{M_1}) - l_n^{M_2}(\hat{\boldsymbol{\theta}}_{ML}^{M_2}) \right) - \log(n)(d_{M_1} - d_{M_2}) \end{aligned}$$

i.e. it is the penalised log-likelihood ratio of model M_1 vs. M_2.

12.2.4 Model complexity and Occams razor

As demonstrated above the averaging over θ in the marginal likelihood has the effect of automatically penalising complex models.

Therefore, when comparing models using marginal likelihood, as in the Bayes factor, a complex model may be ranked below simpler models. In contrast, when selecting a model by maximum likelihood directly, without averaging, the model with the highest number of parameters always wins over simpler models.

Thus, the penalisation implicit in the marginal likelihood is very much desired as it prevents the overfitting of maximum likelihood. **The principle of preferring a less complex model is called "Occam's razor"**, and it is a natural propery of the Bayes factor.

Note that when comparing models a simpler model is often preferably over a more complex model, because the simpler model is typically better suited to both explaining the currently observed data as well as future data, whereas a complex model will only excel in fitting the current data but will then perform poorly in prediction.

Chapter 13

False discovery rates

13.1 General setup

13.1.1 Overview

In this chapter we introduce False Discovery Rates (FDR) as a Bayesian method to distinguish a null model from an alternative model. This is closely linked with classical frequentist multiple testing procedures.

13.1.2 Choosing between H_0 and H_A

We consider two models:

H_0 : null model, with density $f_0(x)$ and distribution $F_0(x)$

H_A : alternative model, with density $f_A(x)$ and distribution $F_A(x)$

Aim: given observations $x_1, \ldots, x_n$ we would like to decide for each x_i whether it belongs to H_0 or H_A.

This is done by a critical decision threshold x_c: if $x_i > x_c$ then x_i is called "significant" and otherwise called "not significant".

In classical statistics one of the the most widely used approach to find the decision threshold is by computing p-values from the x_i (this uses only the null model but not the alternative model), and then thresholding the p-values a a certain level (say 5%). If n is large then often the test is modified by adjusting the p-values or the threshold (e.g. if Bonferroni correction).

Note that this procedure ignores any information we may have about the alternative model!

13.1.3 True and false positives and negatives

For any decision threshold x_c we can distinguish the following errors:

- False positives (FP), "false alarm", type I error: x_i belongs to null but is called "significant"

- False negative (FN), "miss", type II error: x_i belongs to alternative, but is called "not significant"

In addition we have:

- True positives (TP), "hits": belongs to alternative and is called "significant"
- True negatives (TN), "correct rejections": belongs to null and is called "not significant"

13.2 Specificity and Sensitivity

From counts of TP, TN, FN, FP we can derive further quantities:

- True Negative Rate TNR, **specificity**: $TNR = \frac{TN}{TN+FP} = 1 - FPR$ with FPR=False Positive Rate = $1 - \alpha_I$
- True Positive Rate TPR, **sensitivity**, **power**, recall: $TNR = \frac{TP}{TP+FN} = 1 - FNR$ with FNR=False negative rate = $1 - \alpha_{II}$
- Accuracy: $ACC = \frac{TP+TN}{TP+TN+FP+FN}$

Another common way to choose the decision threshold x_d in classical statistics is to balance sensitivity/power vs. specificity (maximising both power and specificity, or equivalently, minimising both false positive and false negative rates). ROC curves plot TPR/sensitivity vs. FPR = 1-specificity.

13.3 FDR and FNDR

It is possible to link the above with the observed counts of TP, FP, TN, FN:

- False Discovery Rate (FDR): $FDR = \frac{FP}{FP+TP}$
- False Nondiscovery Rate (FNDR): $FNDR = \frac{FN}{TN+FN}$
- Positive predictive value (PPV), True Discovery Rate (TDR), precision: $PPV = \frac{TP}{FP+TP} = 1 - FDR$
- Negative predictive value (NPV): $NPV = \frac{TN}{TN+FN} = 1 - FNDR$

In order to choose the decision threshold it is natural to balance FDR and FDNR (or PPV and NPV), by minimising both FDR and FNDR or maximising both PPV and NPV.

In machine learning it is common to use "precision-recall plots" that plot precision (=PPV, TDR) vs. recall (=power, sensitivity).

13.4 Bayesian perspective

13.4.1 Two component mixture model

In the Bayesian perspective the problem of choosing the decision threshold is related to computing the posterior probability

$$\Pr(H_0|x_i),$$

i.e. probability of the null model given the observation x_i, or equivalently computing

$$\Pr(H_A|x_i) = 1 - \Pr(H_0|x_i)$$

the probability of the alternative model given the observation x_i.

This is done by assuming a mixture model

$$f(x) = \pi_0 f_0(x) + (1 - \pi_0) f_A(x)$$

where $\pi_0 = \Pr(H_0)$ is the prior probability of H_0 and. $\pi_A = 1 - \pi_0 = \Pr(H_A)$ the prior probabiltiy of H_A.

Note that the weights π_0 can in fact be estimated from the observations by fitting the mixture distribution to the observations $x_1, \ldots, x_n$ (which implies that this yields a form of empirical Bayes method).

13.4.2 Local FDR

The posterior probability of the null model given a data point is then given by

$$\Pr(H_0|x_i) = \frac{\pi_0 f_0(x_i)}{f(x_i)} = LFDR(x_i)$$

This quantity is also known as the **local FDR** or **local False Discovery Rate**.

In the given one-sided setup the local FDR is large (close to 1) for small x, and will become close to 0 for large x. A common decision rule is given by thresholding local false discovery rates: if $LFDR(x_i) < 0.1$ the x_i is called significant.

13.4.3 q-values

In correspondence to p-values one can also define tail-area based false discovery rates:

$$Fdr(x_i) = \Pr(H_0|X > x_i) = \frac{\pi_0 F_0(x_i)}{F(x_i)}$$

These are called **q-values**, or simply **False Discovery Rates (FDR)**. Intriguingly, these also have a frequentist interpretation as adjusted p-values (using a Benjamini-Hochberg adjustment procedure).

13.5 Software

There are a number of R packages to compute (local) FDR values:

For example:

- locfdr
- qvalue
- fdrtool

and many more.

Using FDR values for screening is especially useful in high-dimensional settings (e.g. when analysing genomic and other high-throughput data).

FDR values have both a Bayesian as well as frequentist interpretation, providing further evidence that good classical statistical methods do have a Bayesian interpretation.

Chapter 14

Optimality properties and summary

14.1 Bayesian statistics in a nutshell

- Bayesian statistics explicitly models the uncertainty about the parameters of interests by probability
- In the light of new evidence (observed data) the uncertainty is updated, i.e. the prior distribution is combined with the likelihood to form the posterior distribution

Example: Beta-Binomial model

- Binomial likelihood
- n observations: x "heads", $n - x$ "tails"
- Frequency $\hat{\theta}_{ML} = \frac{x}{n}$
- Beta prior $\theta \sim Beta(\alpha_0, \beta_0)$ with mean $\theta_0 = \frac{\alpha_0}{m}$ and $m = \alpha_0 + \beta_0$
- Beta posterior $\theta|x, n \sim Beta(\alpha_1, \beta_1)$ with mean $\theta_1 = \frac{\alpha_1}{\alpha_1+\beta_1}$ and $\alpha_1 = \alpha_0 + x$ and $\beta_1 = \beta_0 + n - x$
- Update of prior mean to posterior mean by shrinkage of MLE:

$$\theta_1 = \lambda\theta_0 + (1 - \lambda)\hat{\theta}_{ML}$$

  with shrinkage intensity $\lambda = \frac{m}{n+m}$
- m can be interpreted as prior sample size

14.1.1 Remarks

- If posterior in same family as prior $\rightarrow$ conjugate prior
- In the exponential family the Bayesian update of the mean is always expressible as linear shrinkage of the MLE
- For sample size $n \rightarrow \infty$ then $\lambda \rightarrow 0$ and $\theta_1 \rightarrow \hat{\theta}_{ML}$ (for large samples posterior mean = maximum likelihood estimator)

- For $n \to 0$ then $\lambda \to 1$ and $\theta_1 \to \hat{\theta}_0$ (if no data is available fall back to prior)
- Note that the Bayesian estimator is biased for finite n by construction (but asymptotically unbiased like the MLE).

14.1.2 Advantages

- adding prior information has regularisation properties. This is very important in more complex models with many parameters, e.g., in estimation of a covariance matrix (to avoid singularity).
- improves small-sample accuracy (e.g. MSE)
- that Bayesian estimators tend to be better than MLE is not surprising - they use the data plus extra information!
- Bayesian credible intervals are conceptually much more simple than frequentist confidence intervals

14.2 Frequentist properties of Bayesian estimators

A Bayesian point estimator (e.g. the posterior mean) can also be assessed by its frequentist properties.

- First, we know that, by construction, the Bayesian estimator $\hat{p}_{\text{Bayes}}$ will be biased for finite n even if the MLE is unbiased (with the bias being the posterior mean in this case).
- Second, intriguingly it turns out that the sampling variance of the Bayes point estimator (not to be confused with the posterior variance!) can be smaller than the variance of the MLE. This depends on the choice of the shrinkage parameter λ that also determines the posterior variance.

As a result, Bayesian estimators may have smaller MSE (=squared bias + variance) than the ML estimator for finite n.

In statistical decision theory this is called the theorem of **admissibility of Bayes rules**. It states that under mild conditions every admissible estimation rule (i.e. one that dominates all other estimators with regard to some expected loss, such as the MSE) is in fact a Bayes estimator with some prior.

Unfortunately, this theorem does not tell which prior is needed to achive optimality, however an optimal estimator with minimum MSE can often be found by tuning λ.

14.3 Specifying the prior — problem or advantage?

In Bayesian statistics the analysist needs to be very explicit about the modeling assumptions:

Model = data generating process (likelihood) + prior uncertainty (prior distribution)

Note that alternative statistical methods can often be interpreted as Bayesian methods assuming a specific *implicit* prior!

For example, likelihood estimation for the Binomial model is equivalent to Bayes estimation using the Beta-Binomial model with a $Beta(0, 0)$ prior (=Haldane prior).
However, when choosing a prior explicitly for this model, interestingly most analysts would rather use a flat prior $Beta(1, 1)$ (=Laplace prior) with implicit sample size $m = 2$ or a transformation-invariant prior $Beta(1/2, 1/2)$ (=Jeffreys prior) with implicit sample size $m = 1$ than the Haldane prior!

$\rightarrow$ be aware about the implicit priors!!

Better to acknowledge that a prior is being used (even if implicit!)

Writing down all your assumptions is enforced by the Bayesian approach.

Specifying a prior is thus best understood as an intrinsic part of model specification. It helps to improve inference and it may only be ignored if there is lots of data.

14.4 Choosing a prior

It is **essential in a Bayesian analysis to specify your prior uncertainty about the model parameters**. Note that this is simply **part of the modeling process**!

Typically, the location of the prior determines the amount of bias, and the precision (inverse variance) of the prior is proportional to the implied sample size of the prior.

As we have seen before for large n the Bayesian estimate converges to the ML estimate, so for large n you may ignore specifying a prior.

However, for small n it is essential that a prior is specified. In non-Bayesian approaches (if interpreted from Bayesian perspective) this prior is still there but it is implicit (e.g. uniform prior for likelihood estimation).

14.4.1 Some guidelines

So the questions remains what are good ways to choose a prior? Two useful ways (among many others) are:

1. Use a weakly informative prior (cf. Gelman). This means that you have a vague idea about the suitable values of the parameter of interest, and you use a corresponding prior (with moderate variance) to model the uncertainty. This acknowledges that there are no uninformative priors and aims to ensure that the prior will not dominate the likelihood.

2. Empirical Bayes methods can often be used to determine one or all of the hyperparameters (i.e. the parameters in the prior). There are several ways to do this, one of them is to tune the shrinkage parameter λ to achieve minimum MSE. We discuss this further below.

In contrast, there also exists many proposals advocating to select so-called "uninformative priors". However, as it is easly shown, there are no true unformative priors, since a prior that looks uninformative (i.e. "flat") in one coordinate system can be informative in another — this is a simple consequence of the rule for transformation of probability densities. Furthermore, often these priors are improper, i.e. are not actually probability distributions. For this (and many other reasons) the search for "uniformative" priors is not just futile but in fact also undesirable (e.g. the prior typically also needs to act as regulariser)!

Instead, **specifying the prior needs to be viewed as part of the modelling process, with specification of the prior as integral as the specification of the likelihood.**

14.4.2 Jeffreys prior

In order to complement the discussion on non-informative priors we now look (briefly) at the proposal by Jeffreys (1946).

Specifically, this prior is constructed from the expected Fisher observation using the log-likelihood function and thus promises automatic construction of objective uninformative priors:

$$p(\boldsymbol{\theta}) \propto \sqrt{\det \boldsymbol{I}^{\text{Fisher}}(\boldsymbol{\theta})}$$

The reasoning underlying this prior is **invariance against transformation of the coordinate system of the parameters**.

For the Beta-Binomial model the Jeffreys prior corresponds to $Beta(\frac{1}{2}, \frac{1}{2})$.

For the Normal-Normal model it corresponds to the flat improper prior $p(\mu) = 1$.

For the Inverse-Gamma-Norma model the Jeffreys prior is the improper prior $p(\sigma^2) = \frac{1}{\sigma^2}$.

This illustrate the main problem with this prior – namely that it often is an improper prior.

Another issue is that Jeffreys priors are usually not conjugate which complicates the update to the posterior. An alternative to Jeffreys prior is the **reference prior** developed by Bernardo (1979).

14.5 Optimality of Bayes inference

The optimality of Bayesian model making use of full model specification (likelihood plus prior) can be shown from a number of different perspectives. Correspondingly, there are many theorems that prove (or at least indicate) this optimality:

1) Richard Cox's theorem: the aim to generalise classic logic inevitably leads to Bayesian inference.

2) Entropy perspective: Bayesian inference is a consequence of minimal information update where new information arrives in form of observations

3) de Finetti's representation theorem: joint distribution of exchangeable sequences can be viewed as posterior distributions computed by Bayes theorem)

4) Frequentist decision theory: all admissible decision rules are Bayes rules! (admissible = always better than all other methods!)

Remark: the above also excludes a few other (somewhat esoteric) suggestions for propagating uncertainty (e.g. Fuzzy Logic, imprecise probabilties, etc).

14.6 Conclusion

Bayesian statistics offers a coherent framework for statistical learning from data, with methods for

- estimation
- testing
- model building

There are a number of theorems that show that "optimal" estimators (defined in various ways) are all Bayesian.

Bayesian statistics is a non-asymptotic theory, it works for any sample size.

Note that many classical (frequentist) procedures may be viewed as *approximations* to Bayesian methods and estimators, so using classical approaches in the correct application domain is perfectly in line with the Bayesian framework.

14.6.1 Current research

For example: connection between Bayesian models and algorithmic models widely used in machine learning (such as neural networks, deep learning, convolutional networks, ensemble methods, XGBoost etc).

Are these models optimal (as in the Bayesian sense)? Can we learn something about highly complex, non-parametric statistical models?

How do we do effective Bayesian learning for these parameter-rich models? Both in terms of computational and statistical efficiency!

Part III

Regression

Chapter 15

Overview over regression modelling

15.1 General setup

- y: **response variable**, also known as **outcome** or **label**
- $x_1, x_2, x_3, \ldots, x_d$: **predictor variables**, also known as **covariates** or **covariables**
- The relationship between the outcomes and the predictor variables is assumed to follow

  $$y = f(x_1, x_2, \ldots, x_d) + \varepsilon$$

  where f is the **regression function** (not a density) and ε represents **noise**.

15.2 Objectives

1. **Understand the relationship** between the response y and the predictor variables x_i by **learning the regression function** f from observed data (training data). The estimated regression function is $\hat{f}$.

2. **Prediction of outcomes**

$$\underbrace{\hat{y}}_{\substack{\text{predicted response} \\ \text{using fitted } \hat{f}}} = \hat{f}(x_1, x_2, \ldots, x_d)$$

If instead of the fitted function $\hat{f}$ the known regression function f is used we denote this by

$$\underbrace{y^\star}_{\substack{\text{predicted response} \\ \text{using known } f}} = f(x_1, x_2, \ldots, x_d)$$

3. **Variable importance**

- which covariates are most relevant in predicting the outcome?
- allows to better understand the data and model
  $\rightarrow$ variable selection (to build simpler model with same predictive capability)

15.3 Regression as a form of supervised learning

Regression modeling is a special case of **supervised learning**.

In supervised learning we make use of labeled data, i.e. $\boldsymbol{x}_i$ has an associated *label* y_i. Thus, the data is consists of pairs $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)$.

The *supervision* part of in supervised learning refers to the fact that the labels are given.

In **regression** typically the label y_i is continuous and called the *response*.

On the other hand, if the label y_i is discrete/categorical then supervised learning is called **classification**.

$$\text{Supervised Learning} \begin{cases} \longrightarrow \text{Discrete } y & \longrightarrow \text{Classification Methods} \\ \longrightarrow \text{Continuous } y & \longrightarrow \text{Regression Methods} \end{cases}$$

Another important type of statistical learning is **unsupervised learning** where labels y are inferred from the data $\boldsymbol{x}$ (this is also known as **clustering**). Furthermore, there is also *semi-supervised learning* with labels only partly known.

Note that there are regression models (e.g. logistic regression) with discrete response that are performing classification, so one may argue that "supervised learning"="generalised regression".

15.4 Various regression models used in statistics

In this course we only study linear multiple regression. However, you should be aware that the linear model is in fact just a special cases of some much more general regression approaches.

General regression model:

$$y = f(x_1, \ldots, x_d) + \text{"noise"}$$

- The function f is estimated nonparametrically - splines - Gaussian processes
- Generalised Additive Models (GAM): - the function f is assumed to be the sum of individual functions $f_i(x_i)$
- Generalised Linear Models (GLM): - f is a transformed linear predictor $h(\sum b_i x_i)$, noise is assumed from exponential family
- Linear Model (LM): - linear predictor $\sum b_i x_i$, normal noise

In R the linear model is implemented in the function lm(), and generalised linear models in the function glm(). Generalised additive models are available in the package “mgcv”.

In the following we focus on the linear regression model with continuous response.

Chapter 16

Linear Regression

16.1 The linear regression model

In this module we assume that f is a linear function:

$$f(x_1, \ldots, x_d) = \beta_0 + \sum_{j=1}^{d} \beta_j x_j = y^\star$$

In vector notation:

$$f(\boldsymbol{x}) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x} = y^\star$$

with $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_d \end{pmatrix}$ and $\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$

Therefore, the linear regression model is

$$\begin{aligned} y &= \beta_0 + \sum_{j=1}^{d} \beta_j x_j + \varepsilon \\ &= \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x} + \varepsilon \\ &= y^\star + \varepsilon \end{aligned}$$

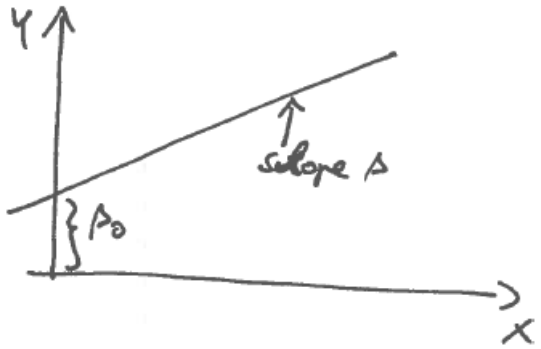
where:

- β_0 is the **intercept**
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^T$ are the **regression coefficients**
- $\boldsymbol{x}$ is the predictor vector

16.2 Interpretation of regression coefficients and intercept

- The regression coefficient β_i corresponds to the slope (first partial derivative) of the regression function in the direction of x_i. In other words, the gradient of $f(\boldsymbol{x})$ are the regression coefficients: $\nabla f(\boldsymbol{x}) = \boldsymbol{\beta}$

- The intercept β_0 is the offset at the origin ($x_1 = x_2 = \ldots = x_d = 0$):



16.3 Different types of linear regression:

- **Simple linear regression**: $y = \beta_0 + \beta x + \epsilon$ (=single predictor)
- **Multiple linear regression**: $y = \beta_0 + \sum_{j=1}^{d} \beta_j x_j + \epsilon$ (= multiple predictor variables)
- **Multivariate regression**: multivariate response $\boldsymbol{y}$

16.4 Distributional assumptions and properties

General assumptions:

In our application we treat y and $x_1, \ldots, x_d$ as observables that can be described by random variables.
$\beta_0, \boldsymbol{\beta}$ are parameters to be inferred from the observations on y and $x_1, \ldots, x_d$.

- Specifically, will we assume that response and predictors have a mean and a (cov)variance:

  i. Response:
  $\mathrm{E}(y) = \mu_y$
  $\mathrm{Var}(y) = \sigma_y^2$
  The **variance of the response** $\mathrm{Var}(y)$ is also called the **total variation** .

  ii. Predictors:
  $\mathrm{E}(x_i) = \mu_{x_i}$ (or $\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\mu_x}$)
  $\mathrm{Var}(x_i) = \sigma_{x_i}^2$ and $\mathrm{Cor}(x_i, x_j) = \rho_{ij}$ (or $\mathrm{Var}(\boldsymbol{x}) = \boldsymbol{\Sigma_x}$)
  The **signal variance** $\mathrm{Var}(y^\star) = \mathrm{Var}(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}) = \boldsymbol{\beta}^T \boldsymbol{\Sigma_x} \boldsymbol{\beta}$ is also called the **explained variation**.

  (Note: see next lecture for discussion and definition of the covariance matrix $\boldsymbol{\Sigma}$).

- The noise ε is also an observable, but typically only indirectly via the difference $\varepsilon = y - y^\star$. We denote the mean and variance of the noise by

$\mathrm{E}(\varepsilon)$ and $\mathrm{Var}(\varepsilon)$.
The **noise variance** $\mathrm{Var}(\varepsilon)$ is also called the **unexplained variation**.

- We assume that y and $\boldsymbol{x}$ are jointly distributed with some correlation $\mathrm{Cor}(y, x_j) = \rho_{y,x_j}$ between each predictor variable x_j and the response y.

Identifiability assumptions:

In a statistical analysis we would like to be able to separate signal ($y^\star$) from noise (ε). To achieve this we require some **distributional assumptions to ensure identifiability** and avoid confounding:

1) **Assumption 1:** ε and $y^\star$ are are independent. This implies $\mathrm{Var}(y) = \mathrm{Var}(y^\star) + \mathrm{Var}(\varepsilon)$, or equivalently $\mathrm{Var}(\varepsilon) = \mathrm{Var}(y) - \mathrm{Var}(y^\star)$.

   Thus, this assumption implies the **decomposition of variance**, i.e. that the **total variation** $\mathrm{Var}(y)$ equals the sum of the **explained variation**$\mathrm{Var}(y^\star)$ and the **unexplained variation**$\mathrm{Var}(\varepsilon)$.

2) **Assumption 2:** $\mathrm{E}(\varepsilon) = 0$. This allows to identify the intercept β_0 and implies $\mathrm{E}(y) = \mathrm{E}(y^\star)$.

Optional assumptions (often but not always):

- The noise ε is normally distributed
- The response y and and the predictor variables x_i are continuous variables
- The response and predictor variables are jointly normally distributed

Further properties:

- As a result of the independence assumption 1) we can only choose two out of the three variances freely:
  i. in a generative perspective we will choose signal variance $\mathrm{Var}(y^\star)$ (or equivalently the variances $\mathrm{Var}(x_j)$) and the noise variance $\mathrm{Var}(\varepsilon)$, then the variance of the response $\mathrm{Var}(y)$ follows.
  ii. in an observational perspective we will observe the variance of the reponse $\mathrm{Var}(y)$ and the variances $\mathrm{Var}(x_j)$, and then the error variance $\mathrm{Var}(\varepsilon)$ follows.
- As we will see later the regression coefficients β_j depend on the correlations between the response y and and the predictor variables x_j. Thus, the choice of regression coefficients implies a specific correlation pattern, and vice vera (in fact, we will use this correlation pattern to infer the regression coefficient from data!).

16.5 Regression in data matrix notation

We can also write the regression in terms of actual observed data (rather than random variables):

Data matrix for the predictors:

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nd} \end{pmatrix}$$

Note the statistics convention: the n rows of $\boldsymbol{X}$ contain the samples, and the d columns contain variables.

Response data vector: $(y_1, \ldots, y_n)^T = \boldsymbol{y}$

Then the regression equation is written in data matrix notation:

$$\underbrace{\boldsymbol{y}}_{n\times 1} = \underbrace{\mathbf{1}_n \beta_0}_{n\times 1} + \underbrace{\boldsymbol{X}}_{n\times d}\underbrace{\boldsymbol{\beta}}_{d\times 1} + \underbrace{\boldsymbol{\varepsilon}}_{\underbrace{n\times 1}_{\text{residuals}}}$$

where $\mathbf{1}_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ is a column vector of length n (size $n \times 1$).

Note that here the regression coefficients are now multiplied *after* the data matrix (compare with the original vector notation where the *transpose* of regression coefficients come *before* the vector of the predictors).

The **observed noise** values (i.e. realisations of ε) are called the **residuals**.

16.6 Centering and vanishing of the intercept β_0

If $\boldsymbol{x}$ and y are centered, i.e. $\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\mu_x} = 0$ and $\mathrm{E}(y) = \mu_y = 0$ then the intercept β_0 disappears:

The regression equation is

$$y = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x} + \varepsilon$$

with $E(\varepsilon)$. Taking the expectation on both sides we get $\mu_y = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{\mu_x}$ and therefore

$$\beta_0 = \mu_y - \boldsymbol{\beta}^T \boldsymbol{\mu_x}$$

This is equal to zero if the means of the response and predictors vanish. Conversely, if we assume that the intercept vanishes ($\beta_0 = 0$) this is only possible for general $\boldsymbol{\beta}$ if both $\boldsymbol{\mu_x} = 0$ and $\mu_y = 0$.

Thus, in the linear model is always possible to transform y and $\boldsymbol{x}$ (or data $\boldsymbol{y}$ and $\boldsymbol{X}$) so that the intercept vanishes!
$\Rightarrow$ we will therefore often set $\beta_0 = 0$.

16.7 Regression objectives for linear model

1. Understand functional relationship: find estimates of intercept ($\hat{\beta}_0$) and regression coefficients ($\hat{\beta}_j$)
2. Prediction:
   - Known coefficients β_0 and $\boldsymbol{\beta}$: $y^\star = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}$

- Estimated coefficients $\hat{\beta}_0$ and $\hat{\beta}$ (note the "hat"!): $\hat{y} = \hat{\beta}_0 + \sum_{j=1}^{d} \hat{\beta}_j x_j = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$

Also find the **corresponding prediction errors!**

3. Variable importance: Which predictors x_j are most relevant?
   $\rightarrow$ test whether $\beta_j = 0$
   $\rightarrow$ find measures of variable importance
   Remark: as we will see β_j or $\hat{\beta}_j$ itself is **not** a measure of importance!

Chapter 17

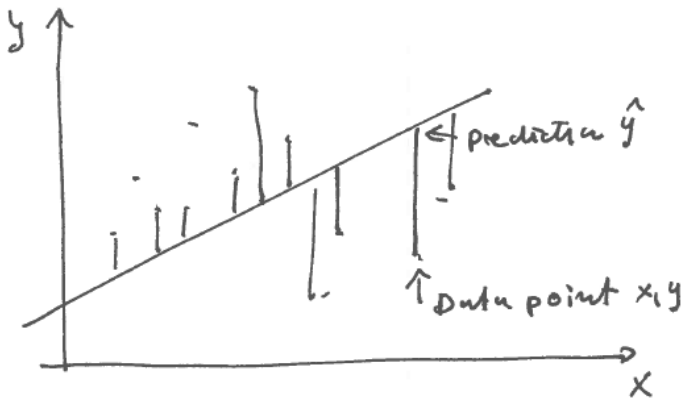
Estimating regression coefficients

In this chapter we discuss various ways to estimate the regression coefficients. First, we discuss estimation by Ordinary Least Squares (OLS) by minimising the residual sum of squares. This yields the famous Gauss estimator. Second, we derive estimates of the regression coefficients using the methods of maximum likelihood assuming normal errors. This also leads to the Gauss estimator. Third, we show that the coefficients in linear regression can written and interpreted in terms of two covariance matrices and that the Gauss estimator of the regression coefficients is a plug-in estimator using the MLEs of these covariance matrices. Furthermore, we show that the (population version) of the Gauss estimator can also be derived by finding the best linear predictor and by conditioning. Finally, we discuss special cases of regression coefficients and their relationship to marginal correlation.

17.1 Ordinary Least Squares (OLS) estimator of regression coefficients

Now we show the classic way (Gauss 1809; Legendre 1805) to **estimate regression coefficients** by the method of **ordinary least squares (OLS)**.

Idea: choose regression coefficients such as to *minimise* the *squared error* between observations and the prediction (=RSS, the Residual Sum of Squares, a function of β)



In data matrix notation (note we assume $\beta_0 = 0$ and thus *centered data* $\boldsymbol{X}$ and $\boldsymbol{y}$):

$$\text{RSS}(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = \sum_{i=1}^{n} \epsilon_i^2$$

$$\widehat{\boldsymbol{\beta}}_{\text{OLS}} = \arg\min \text{RSS}(\boldsymbol{\beta})$$

$$\text{RSS}(\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}$$

Gradient:

$$\nabla\text{RSS}(\boldsymbol{\beta}) = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}$$

$$\nabla\text{RSS}(\widehat{\boldsymbol{\beta}}) = 0 \longrightarrow \boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}$$

$$\Longrightarrow \widehat{\boldsymbol{\beta}}_{\text{OLS}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Note the similarities in the procedure to maximum likelihood (ML) estimation (with minimisation instead of maximisation)! In fact, as we see nextm this is not by chance as OLS *is* indeed a special case of ML! This also implies that OLS is generally a good method — but only if sample size n is large!

The above Gauss' estimator is fundamental in statistics so it is worthwile to memorise it!

17.2 Maximum likelihood estimation of regression coefficients

We now show how to estimate regression coefficients using the method of maximum likelihood. This is a second method to derive $\hat{\boldsymbol{\beta}}$.

We recall the basic regression equation

$$y = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x} + \epsilon$$

with $\mathrm{E}(\varepsilon) = 0$ and observed data $y_1, \ldots, y_n$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. The intercept is identified as

$$\beta_0 = \mu_y - \boldsymbol{\beta}^T \boldsymbol{\mu}_{\boldsymbol{x}}$$

so that we can solve for the noise variable

$$\varepsilon = (y - \mu_y) - \boldsymbol{\beta}^T (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}})$$

Assuming joint (multivariate) normality for the response y and $\boldsymbol{x}$ we get as the MLEs for the respective means and (co)variances:

- $\hat{\mu}_y = \widehat{\mathrm{E}(y)} = \frac{1}{n} \sum_{i=1}^n y_i$
- $\hat{\sigma}_y^2 = \widehat{\mathrm{Var}(y)} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\mu}_y)^2$
- $\hat{\boldsymbol{\mu}}_{\boldsymbol{x}} = \hat{\mathrm{E}}(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_i$
- $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}} = \widehat{\mathrm{Var}(\boldsymbol{x})} = \frac{1}{n} \sum_{i=1}^n (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})^T$
- $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} = \widehat{\mathrm{Cov}(\boldsymbol{x}, y)} = \frac{1}{n} \sum_{i=1}^n (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})(y_i - \hat{\mu}_y)$

The noise $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ is normally distributed with mean 0 and variance $\mathrm{Var}(\varepsilon) = \sigma_\varepsilon^2$ which leads to the normal log-likelihood function:

$$\begin{aligned}
\log L &= -\frac{n}{2} \log \sigma_\varepsilon^2 - \frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^n \epsilon_i^2 \\
&= -\frac{n}{2} \log \sigma_\varepsilon^2 - \frac{1}{2\sigma_\varepsilon^2} \underbrace{\sum_{i=1}^n \Big((y_i - \hat{\mu}_y) - \boldsymbol{\beta}^T (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}}) \Big)^2}_{= \mathrm{RSS}(\boldsymbol{\beta})}
\end{aligned}$$

We now only need to maximise the log-likelihood to obtain MLEs of σ_ε^2 and $\boldsymbol{\beta}$!

Note that the residual sum of squares (RSS) appears in the log-likelihood function (with a minus sign), which implies that ML assuming normal distribution will recover the OLS estimator for the regression coefficients! So OLS is a special case of ML !

17.2.1 Detailed derivation of the MLEs

The gradient with regard to $\boldsymbol{\beta}$ is

$$\begin{aligned}
\nabla_{\boldsymbol{\beta}} \log L &= \frac{1}{\sigma_\varepsilon^2} \sum_{i=1}^n \Big((\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})(y_i - \hat{\mu}_y) - (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})^T \boldsymbol{\beta} \Big) \\
&= \frac{n}{\sigma_\varepsilon^2} \left(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} - \hat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}} \boldsymbol{\beta} \right)
\end{aligned}$$

Setting this equal to zero yields the Gauss estimator

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}}$$

By plugin we the get the MLE of β_0 as

$$\hat{\beta}_0 = \hat{\mu}_y - \hat{\boldsymbol{\beta}}^T \hat{\boldsymbol{\mu}}_{\boldsymbol{x}}$$

Taking the derivate of $\log L$ with regard to σ_ε^2 results in

$$\frac{\partial}{\partial \sigma_\varepsilon^2} \log L = -\frac{n}{2\sigma_\varepsilon^2} + \frac{1}{2\sigma_\varepsilon^4} RSS(\beta)$$

which leads to the MLE

$$\hat{\sigma^2}_\varepsilon = \frac{RSS(\hat{\boldsymbol{\beta}})}{n} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

with $\hat{y}_i = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$.

Note that the MLE $\widehat{\sigma^2}_\varepsilon$ is a biased estimate of σ_ε^2. The unbiased estimate is $\frac{1}{n-d-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$, where d is the dimension of $\boldsymbol{\beta}$ (i.e. the number of predictors).

17.2.2 Asymptotics

The advantage of using maximum likelihood is that we also get the (asympotic) variance associated with each estimator and typically can also assume asymptotic normality.

Specifically, for $\hat{\boldsymbol{\beta}}$ we get via the observed Fisher information at the MLE an asymptotic estimator of its variance

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}) = \frac{1}{n} \hat{\sigma^2}_\varepsilon \hat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}^{-1}$$

Similarly, for $\hat{\beta}_0$ we have

$$\widehat{\mathrm{Var}}(\widehat{\beta}_0) = \frac{1}{n} \hat{\sigma^2}_\varepsilon (1 + \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}^{-1} \hat{\boldsymbol{\mu}})$$

For finite sample size n with known $\mathrm{Var}(\varepsilon)$ one can show that the variances are

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \frac{1}{n} \sigma_\varepsilon^2 \hat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}^{-1}$$

and

$$\mathrm{Var}(\widehat{\beta}_0) = \frac{1}{n} \sigma_\varepsilon^2 (1 + \hat{\boldsymbol{\mu}}_{\boldsymbol{x}}^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}^{-1} \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})$$

and that the regression coefficients and the intercept are normally distributed according to

$$\widehat{\boldsymbol{\beta}} \sim N_d(\boldsymbol{\beta}, \mathrm{Var}(\widehat{\boldsymbol{\beta}}))$$

and

$$\widehat{\beta}_0 \sim N(\beta_0, \mathrm{Var}(\widehat{\beta}_0))$$

We may use this to test whether whether $\beta_j = 0$ and $\beta_0 = 0$.

17.3 Covariance plug-in estimator of regression coefficients

We now try to understand regression coefficients in terms of covariances (thus obtaining a third way to compute and estimate them).

We recall that the Gauss regression coefficients are given by

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

where $\boldsymbol{X}$ is the $n \times d$ data matrix (in statistics convention)

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nd} \end{pmatrix}$$

Note that we assume that the data matrix $\boldsymbol{X}$ is centered (i.e. column sums $\boldsymbol{X}^T\mathbf{1}_n = \mathbf{0}$ are zero).

Likewise $\boldsymbol{y} = (y_1, \dots, y_n)^T$ is the response data vector (also centered with $\boldsymbol{y}^T\mathbf{1}_n = 0$).

Noting that

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}} = \frac{1}{n}(\boldsymbol{X}^T\boldsymbol{X})$$

is the MLE of covariance matrix among $\boldsymbol{x}$ and

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} = \frac{1}{n}(\boldsymbol{X}^T\boldsymbol{y})$$

is the MLE of the covariance between $\boldsymbol{x}$ and y we see that the OLS estimate of the regression coefficients can be expressed as

$$\widehat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}\right)^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}}$$

We can write down a population version (with no hats!):

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\boldsymbol{xx}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{xy}}$$

Thus, OLS regression coefficients can be interpreted as plugin estimator using MLEs of covariances! In fact, we may also use the unbiased estimates since the scale factor ($1/n$ or $1/(n-1)$) cancels out so it does not matter which one you use!

17.3.1 Importance of positive definiteness of estimated covariance matrix

Note that $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}$ is inverted in $\widehat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}\right)^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}}$.

- Hence, the estimate $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}$ needs to be positive definite!

- But $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}^{\text{MLE}}$ is only positive definite if $n > d$!

Therefore we can use the ML estimate (empirical estimator) only for large $n > d$, otherwise we need to employ a different (regularised) estimation approach (e.g. Bayes or a penalised ML)!

Remark: writing $\hat{\boldsymbol{\beta}}$ explicitly based on covariance estimates has the advantage that we can construct plug-in estimators of regression coefficient based on regularised covariance estimators that improve over ML for small sample size. This leads to the so-called SCOUT method (=covariance-regularized regression by Witten and Tibshirani, 2008).

17.4 Best linear predictor

The **best linear predictor** is a fourth way to arrive at the linear model. This is closely related to OLS and minimising squared residual error.

Without assuming normality the above multiple regression model can be shown to be optimal linear predictor under the minimum mean squared prediction error:

Assumptions:

- y and $\boldsymbol{x}$ are random variables
- we construct a new variable (the linear predictor) $y^{\star\star} = b_0 + \boldsymbol{b}^T\boldsymbol{x}$ to optimally approximate y

Aim:

- choose b_0 and $\boldsymbol{b}$ such to minimize the mean squared prediction error $\mathrm{E}((y - y^{\star\star})^2)$

17.4.1 Result:

The mean squared prediction error $MSPE$ in dependence of $(b_0, \boldsymbol{b})$ is

$$\begin{aligned}
\mathrm{E}((y - y^{\star\star})^2) &= \mathrm{Var}(y - y^{\star\star}) + \mathrm{E}(y - y^{\star\star})^2 \\
&= \mathrm{Var}(y - b_0 - \boldsymbol{b}^T\boldsymbol{x}) + (\mathrm{E}(y) - b_0 - \boldsymbol{b}^T\mathrm{E}(\boldsymbol{x}))^2 \\
&= \sigma_y^2 + \mathrm{Var}(\boldsymbol{b}^T\boldsymbol{x}) + 2\,\mathrm{Cov}(y, -\boldsymbol{b}^T\boldsymbol{x}) + (\mu_y - b_0 - \boldsymbol{b}^T\boldsymbol{\mu}_{\boldsymbol{x}})^2 \\
&= \sigma_y^2 + \boldsymbol{b}^T\boldsymbol{\Sigma}_{\boldsymbol{xx}}\boldsymbol{b} - 2\,\boldsymbol{b}^T\boldsymbol{\Sigma}_{\boldsymbol{xy}} + (\mu_y - b_0 - \boldsymbol{b}^T\boldsymbol{\mu}_{\boldsymbol{x}})^2 \\
&= MSPE(b_0, \boldsymbol{b})
\end{aligned}$$

We look for

$$(\beta_0, \boldsymbol{\beta}) = \underset{b_0, \boldsymbol{b}}{\arg\min}\ MSPE(b_0, \boldsymbol{b})$$

In order to find the minimum we compute the gradient with regard to $(b_0, \boldsymbol{b})$

$$\nabla MSPE = \begin{pmatrix} -2(\mu_y - b_0 - \boldsymbol{b}^T\boldsymbol{\mu}_{\boldsymbol{x}}) \\ 2\,\boldsymbol{\Sigma}_{\boldsymbol{xx}}\boldsymbol{b} - 2\,\boldsymbol{\Sigma}_{\boldsymbol{xy}} - 2\boldsymbol{\mu}_{\boldsymbol{x}}(\mu_y - b_0 - \boldsymbol{b}^T\boldsymbol{\mu}_{\boldsymbol{x}}) \end{pmatrix}$$

and setting this equal to zero yields

$$\begin{pmatrix} \beta_0 \\ \boldsymbol{\beta} \end{pmatrix} = \begin{pmatrix} \mu_y - \boldsymbol{\beta}^T \boldsymbol{\mu}_x \\ \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\Sigma}_{xy} \end{pmatrix}$$

Thus, the optimal values for b_0 and $\boldsymbol{b}$ in the best linear predictor correspond to the previously derived coefficients β_0 and $\boldsymbol{\beta}$!

17.4.2 Irreducible Error

The minimum achieved MSPE (=**irreducible error**) is

$$MSPE(\beta_0, \boldsymbol{\beta}) = \sigma_y^2 - \boldsymbol{\beta}^T \boldsymbol{\Sigma}_{xx} \boldsymbol{\beta} = \sigma_y^2 - \boldsymbol{\Sigma}_{yx} \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\Sigma}_{xy}$$

With the abbreviation $\Omega^2 = \boldsymbol{P}_{yx} \boldsymbol{P}_{xx}^{-1} \boldsymbol{P}_{xy} = \sigma_y^{-2} \boldsymbol{\Sigma}_{yx} \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\Sigma}_{xy}$ we can simplify this to

$$MSPE(\beta_0, \boldsymbol{\beta}) = \sigma_y^2 (1 - \Omega^2) = \mathrm{Var}(\varepsilon)$$

Writing $b_0 = \beta_0 + \Delta_0$ and $\boldsymbol{b} = \boldsymbol{\beta} + \boldsymbol{\Delta}$ it is easy to see that the mean squared predictive error is a quadratic function around the minimum:

$$MSPE(\beta_0 + \Delta_0, \boldsymbol{\beta} + \boldsymbol{\Delta}) = \mathrm{Var}(\varepsilon) + \Delta_0^2 + \boldsymbol{\Delta}^T \boldsymbol{\Sigma}_{xx} \boldsymbol{\Delta}$$

Note that usually $y^\star = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}$ does not perfectly approximate y so there *will* be an irreducible error (= noise variance)

$$\mathrm{Var}(\varepsilon) = \sigma_y^2 (1 - \Omega^2) > 0$$

which implies $\Omega^2 < 1$.

The quantity Ω^2 has a further interpretation of the population version of as the squared multiple correlation coefficient between the response and the predictors and plays a vital role in decomposition of variance, as discussed later.

17.5 Regression by conditioning

Conditioning is a fifth way to arrive at the linear model. This is also the most general way and can be used to derive many other regression models (not just the simple linear model).

17.5.1 General idea:

- two random variables y (response, scalar) and $\boldsymbol{x}$ (predictor variables, vector)
- we assume that y and $\boldsymbol{x}$ have a joint distribution $F_{y,\boldsymbol{x}}$
- compute *conditional* random variable $y|\boldsymbol{x}$ and the corresponding distribution $F_{y|\boldsymbol{x}}$

17.5.2 Multivariate normal assumption

Now we assume that y and $\boldsymbol{x}$ are (jointly) multivariate normal. Then the conditional distribution $F_{y|\boldsymbol{x}}$ is a univariate normal with the following moments (you can verify this by looking up the general conditional multivariate normal distribution):

a) Conditional expectation:

$$\mathrm{E}(y|\boldsymbol{x}) = y^\star = \beta_0 + \boldsymbol{\beta}^T\boldsymbol{x}$$

with coefficients $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$ and intercept $\beta_0 = \mu_y - \boldsymbol{\beta}^T\boldsymbol{\mu}_{\boldsymbol{x}}$.

Note that as $y^\star$ depends on $\boldsymbol{x}$ it is a random variable itself with mean

$$\mathrm{E}(y^\star) = \beta_0 + \boldsymbol{\beta}^T\boldsymbol{\mu}_{\boldsymbol{x}} = \mu_y$$

and variance

$$\mathrm{Var}(y^\star) = \mathrm{Var}(\mathrm{E}(y|\boldsymbol{x})) = \boldsymbol{\beta}^T\boldsymbol{\Sigma}_{xx}\boldsymbol{\beta} = \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy} = \sigma_y^2\boldsymbol{P}_{yx}\boldsymbol{P}_{xx}^{-1}\boldsymbol{P}_{xy} = \sigma_y^2\Omega^2$$

.

b) Conditional variance:

$$\mathrm{Var}(y|\boldsymbol{x}) = \sigma_y^2 - \boldsymbol{\beta}^T\boldsymbol{\Sigma}_{xx}\boldsymbol{\beta} = \sigma_y^2 - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy} = \sigma_y^2(1-\Omega^2)$$

Note this is a constant so $\mathrm{E}(\mathrm{Var}(y|\boldsymbol{x})) = \sigma_y^2(1-\Omega^2)$ as well.

17.6 Standardised regression coefficients and relationship to correlation

First we note that we can decompose regression coefficients into the product of marginal correlations and correlations among predictors.

Using the variance-correlation decompositions $\boldsymbol{\Sigma}_{xx} = \boldsymbol{V}_{\boldsymbol{x}}^{1/2}\boldsymbol{P}_{xx}\boldsymbol{V}_{\boldsymbol{x}}^{1/2}$ and $\boldsymbol{\Sigma}_{xy} = \boldsymbol{V}_{\boldsymbol{x}}^{1/2}\boldsymbol{P}_{xy}\sigma_y$ we rewrite the regression coefficients as

$$\boldsymbol{\beta} = \underbrace{\boldsymbol{V}_{\boldsymbol{x}}^{-1/2}}_{\text{(inverse) scale of } x_i} \quad \underbrace{\boldsymbol{P}_{xx}^{-1}}_{\text{(inverse) correlation among predictors}} \quad \underbrace{\boldsymbol{P}_{xy}}_{\text{marginal correlations}} \quad \underbrace{\sigma_y}_{\text{scale of } y}$$

Thus the regression coefficients $\boldsymbol{\beta}$ contain the scale of the variables, and take into account the correlations among the predictors ($\boldsymbol{P}_{xx}$) in addition to the marginal correlations between the response y and the predictors x_i ($\boldsymbol{P}_{xy}$).

This decomposition allows to understand a number special cases when the regression coefficient simplify further:

a) If the response and the predictors are standardised to have variance one, i.e. $\mathrm{Var}(y) = 1$ and $\mathrm{Var}(x_i)$, then $\boldsymbol{\beta}$ becomes equal to the **standardised regression coefficients**

$$\boldsymbol{\beta}_{\text{std}} = \boldsymbol{P}_{xx}^{-1}\boldsymbol{P}_{xy}$$

Note that standardised regression coefficients do not make use of variances and and thus are scale-independent.

b) If there is no correlation among the predictors , i.e. $\boldsymbol{P_{xx}} = \boldsymbol{I}$ the the regression coefficients reduce to

$$\boldsymbol{\beta} = \boldsymbol{V_x^{-1}}\boldsymbol{\Sigma}_{xy}$$

where $\boldsymbol{V_x}$ is a diagonal matrix containing the variances oft the predictors. This is also called **marginal regression**. Note that the inversion of $\boldsymbol{V_x}$ is trival since you only need to invert each diagonal element individually.

c) If both a) and b) apply simultaneously (i.e. there is no correlation among predictors and response and predictors and predictors are standardised) then the regression coefficients simplify even further to

$$\boldsymbol{\beta} = \boldsymbol{P}_{xy}$$

Thus, in this very special case the regression coefficients are identical to the correlations between the response and the predictors!

Chapter 18

Squared multiple correlation and variance decomposition in linear regression

In this chapter we first introduce the (squared) multiple correlation and the multiple and adjusted R^2 coefficients as estimators. Subsequently we discuss variance decomposition.

18.1 Squared multiple correlation Ω^2 and the R^2 coefficient

In the previous chapter we encountered the following quantity:

$$\Omega^2 = \boldsymbol{P}_{yx}\boldsymbol{P}_{xx}^{-1}\boldsymbol{P}_{xy} = \sigma_y^{-2}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$$

With $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$ and $\beta_0 = \mu_y - \boldsymbol{\beta}^T\boldsymbol{\mu}_{\boldsymbol{x}}$ it is straightforward to verify the following:

- the cross-covariance between y and $y^\star$ is $\text{Cov}(y, y^\star) = \boldsymbol{\Sigma}_{yx}\boldsymbol{\beta} = \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy} = \sigma_y^2\boldsymbol{P}_{yx}\boldsymbol{P}_{xx}^{-1}\boldsymbol{P}_{xy} = \sigma_y^2\Omega^2$
- the (signal) variance of $y^\star$ is $\text{Var}(y^\star) = \boldsymbol{\beta}^T\boldsymbol{\Sigma}_{xx}\boldsymbol{\beta} = \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy} = \sigma_y^2\boldsymbol{P}_{yx}\boldsymbol{P}_{xx}^{-1}\boldsymbol{P}_{xy} = \sigma_y^2\Omega^2$.

hence the correlation $\text{Cor}(y, y^\star) = \frac{\text{Cov}(y,y^\star)}{\text{SD}(y)\text{SD}(y^\star)} = \Omega$ with $\Omega \geq 0$.

This helps to understand the Ω and Ω^2 coefficients:

- Ω is the linear correlation between the response (y) and prediction $y^\star$.
- Ω^2 is called the **squared multiple correlation** between the scalar y and the vector $\boldsymbol{x}$.

- Note that if we only have one predictor (if x is a scalar) then $\boldsymbol{P}_{xx} = 1$ and $\boldsymbol{P}_{yx} = \rho_{yx}$ so the multiple squared correlation coefficient reduces to squared correlation $\Omega^2 = \rho_{yx}^2$ between y and x.

18.1.1 Estimation of Ω^2 and the multiple R^2 coefficient

The multiple squared correlation coefficient Ω^2 can be estimated by plug-in of empirical estimates for the corresponding correlation matrices:

$$R^2 = \hat{\boldsymbol{P}}_{yx}\hat{\boldsymbol{P}}_{xx}^{-1}\hat{\boldsymbol{P}}_{xy} = \hat{\sigma}_y^{-2}\hat{\boldsymbol{\Sigma}}_{yx}\hat{\boldsymbol{\Sigma}}_{xx}^{-1}\hat{\boldsymbol{\Sigma}}_{xy}$$

This estimator of Ω^2 is called the **multiple R^2 coefficient**.

Note that it does not matter whether the scale factor $1/n$ or $1/(n-1)$ (or something else) is used in estimating the (co)variances since that factor will cancel out when standardising the covariance matrix! So for estimating the correlations it does not matter whether the ML or the unbiased estimator is used.

Above we have seen that Ω^2 is directly linked with the noise variance via

$$\mathrm{Var}(\varepsilon) = \sigma_y^2(1-\Omega^2)\,.$$

An **unbiased estimate** of the noise variance $\mathrm{Var}(\varepsilon)$ (also called **residual variance**) can be computed from the residual sum of squares $RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ and the **degree of freedom** $df = n - d - 1$ by

$$\widehat{\mathrm{Var}}(\varepsilon)_{UB} = \frac{RSS}{df}$$

We can also estimate this from estimates of σ_y^2 and Ω^2. However, with s_y^2 an unbiased estimate of σ_y^2 (with standardisation factor $\frac{1}{n-1}$) a correction factor $\kappa = \frac{n-1}{df}$ is needed to recover the unbiased residual variance:

$$\widehat{\mathrm{Var}}(\varepsilon)_{UB} = s_y^2\,(1-R^2)\,\kappa$$

Setting $(1-R^2)\,\kappa = (1-R^2_{\mathrm{adj}})$ yields the **adjusted multiple R^2 coefficient**

$$R^2_{\mathrm{adj}} = 1 - (1-R^2)\,\kappa$$

so that the unbiased residual variance can be written as

$$\widehat{\mathrm{Var}}(\varepsilon)_{UB} = s_y^2\,(1-R^2_{\mathrm{adj}})$$

18.1.2 R output

In R the command lm() fits the linear regression model.

In addition to the regression cofficients (and derived quantities) the R function lm() also lists

- the multiple R-squared R^2,

- the adjusted R-squared R^2_{adj},
- the degrees of freedom df and
- the residual standard error $\sqrt{\widehat{\text{Var}}(\varepsilon)_{UB}}$ (computed from the unbiased variance estimate).

See also Worksheet 8.

18.2 Variance decomposition in regression

The squared multiple correlation coefficient is useful also because it plays an important role in the decomposition of the total variance:

- total variance: $\text{Var}(y) = \sigma_y^2$
- unexplained variance (irreducible error): $\sigma_y^2(1 - \Omega^2) = \text{Var}(\varepsilon)$
- the explained variance is the complement: $\sigma_y^2\Omega^2 = \text{Var}(y^\star)$

In summary:

$$\text{Var}(y) = \text{Var}(y^\star) + \text{Var}(\varepsilon)$$

becomes

$$\underbrace{\sigma_y^2}_{\text{total variance}} = \underbrace{\sigma_y^2\Omega^2}_{\text{explained variance}} + \underbrace{\sigma_y^2(1-\Omega^2)}_{\text{unexplained variance}}$$

The unexplained variance measures the fit after introducing predictors into the model (smaller means better fit). The total variance measures the fit of the model without any predictors. The explained variance is the difference between total and unexplained variance, it indicates the increase in model fit due to the predictors.

18.2.1 Law of total variance and variance decomposition

The **law of total variance** is

$$\underbrace{\text{Var}(y)}_{\text{total variance}} = \underbrace{\text{Var}(\text{E}(y|\boldsymbol{x}))}_{\text{explained variance}} + \underbrace{\text{E}(\text{Var}(y|\boldsymbol{x}))}_{\text{unexplained variance}}$$

provides a very general decomposition in explained and unexplained parts of the variance that is valid regardless of the form of the distributions $F_{y,\boldsymbol{x}}$ and $F_{y|\boldsymbol{x}}$.

In regression it conncects variance decomposition and conditioning. If you plug-in the conditional expections for the multivariate normal model (cf. previous chapter) we recover

$$\underbrace{\sigma_y^2}_{\text{total variance}} = \underbrace{\sigma_y^2\Omega^2}_{\text{explained variance}} + \underbrace{\sigma_y^2(1-\Omega^2)}_{\text{unexplained variance}}$$

18.2.2 Related quantities

Using the above three quantities (total variance, explained variance, and unexplained variance) we can construct a number of scores:

1) **coefficient of determination**, **squared multiple correlation**: $\frac{\text{explained var}}{\text{total var}} = \frac{\sigma_y^2\Omega^2}{\sigma_y^2} = \Omega^2$
(range 0 to 1, with 1 indicating perfect fit)

2) **coefficient of non-determination**, **coefficient of alienation**: $\frac{\text{unexplained var}}{\text{total var}} = \frac{\sigma_y^2(1-\Omega^2)}{\sigma_y^2} = 1-\Omega^2 = \alpha$
(range 0 to 1, with 0 indicating perfect fit)

3) **inverse alienation coefficient**: $\alpha^{-1} = \frac{\text{total var}}{\text{unexplained var}} = \frac{1}{1-\Omega^2} = 1 + \frac{\text{explained var}}{\text{unexplained var}}$
(range 1 to ∞, with ∞ indicating perfect fit)

4) F **score**, t^2 **score**: $n\frac{\text{explained var}}{\text{unexplained var}} = n\frac{\sigma_y^2\Omega^2}{\sigma_y^2(1-\Omega^2)} = n\frac{\Omega^2}{1-\Omega^2} = \mathcal{F} = \tau^2$
(range 0 to ∞, with ∞ indicating perfect fit)

Note that $\mathcal{F}$ and τ^2 scores (=population versions of F and t^2 statistics) by definition scale with sample size n, and that $\Omega^2 = \frac{\tau^2}{\tau^2+n} = \frac{\mathcal{F}}{\mathcal{F}+n}$ links squared correlation with squared t-scores and F-scores.

18.3 Sample version of variance decomposition

If Ω^2 and σ_y^2 are replaced by their MLEs this can be written in a sample version as follows using data points y_i, predictions $\hat{y}_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$

$$\underbrace{\sum_{i=1}^n (y_i-\bar{y})^2}_{\text{total sum of squares (TSS)}} = \underbrace{\sum_{i=1}^n (\hat{y}_i-\bar{y})^2}_{\text{explained sum of squares (ESS)}} + \underbrace{\sum_{i=1}^n (y_i-\hat{y}_i)^2}_{\text{residual sum of squares (RSS)}}$$

Note that TSS, ESS and RSS all scale with n. Using data vector notation the sample-based variance decomposition can be written in form of the Pythagorean theorem:

$$\underbrace{||\boldsymbol{y}-\bar{y}\mathbf{1}\,||^2}_{\text{total sum of squares (TSS)}} = \underbrace{||\hat{\boldsymbol{y}}-\bar{y}\mathbf{1}||^2}_{\text{explained sum of squares (ESS)}} + \underbrace{||\boldsymbol{y}-\hat{\boldsymbol{y}}||^2}_{\text{residual sum of squares (RSS)}}$$

18.3.1 Geometric interpretation of regression as orthogonal projection:

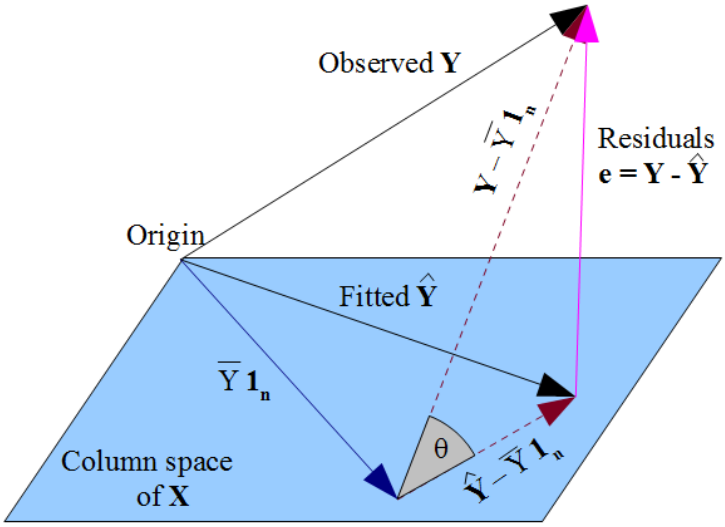
The above equation can be further simplified to

$$||\boldsymbol{y}||^2 = ||\hat{\boldsymbol{y}}||^2 + \underbrace{||\boldsymbol{y} - \hat{\boldsymbol{y}}||^2}_{\text{RSS}}$$

Geometrically speaking, this implies $\hat{\boldsymbol{y}}$ is an orthogonal projection of $\boldsymbol{y}$, since the residuals $\boldsymbol{y} - \hat{\boldsymbol{y}}$ and the predictions $\hat{\boldsymbol{y}}$ are orthogonal (by construction!).

This also valid for the centered versions of the vectors, i.e. $\hat{\boldsymbol{y}} - \bar{y}\mathbf{1}_n$ is an orthogonal projection of $\boldsymbol{y} - \bar{y}\mathbf{1}_n$ (see Figure).

Also note that the angle θ between the two centered vectors is directly related to the (estimated) multiple correlation, with $R = \cos(\theta) = \frac{||\hat{\boldsymbol{y}} - \bar{y}\mathbf{1}_n||}{||\boldsymbol{y} - \bar{y}\mathbf{1}_n||}$, or $R^2 = \cos(\theta)^2 = \frac{||\hat{\boldsymbol{y}} - \bar{y}\mathbf{1}_n||^2}{||\boldsymbol{y} - \bar{y}\mathbf{1}_n||^2} = \frac{\text{ESS}}{\text{TSS}}$.



Source of Figure: Stack Exchange

Chapter 19

Prediction and variable selection

In this chapter we discuss how to compute (lower bounds) of the prediction error and how to select variables relevant for prediction

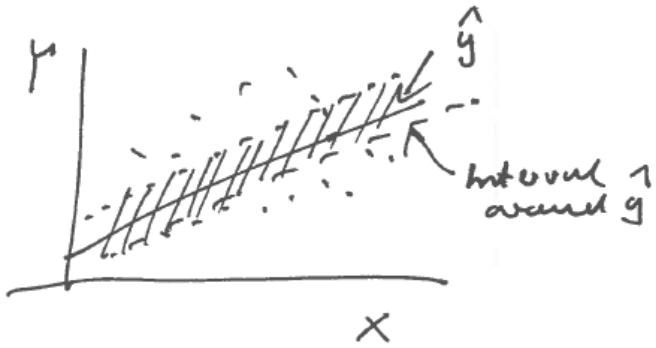
19.1 Prediction and prediction intervals

Learning the regression function from (training) data is only the first step in application of regression models.

The next step is to actually make **prediction** of future outcomes y^{test} given test data $\boldsymbol{x}^{\text{test}}$:

$$y^{\text{test}} = \hat{y}(\boldsymbol{x}^{\text{test}}) = \hat{f}_{\hat{\beta}_0,\hat{\boldsymbol{\beta}}}(\boldsymbol{x}^{\text{test}})$$

Note that y^{test} is a point estimator. Is it possible also to construct a corresponding interval estimate?



The answer is yes, and leads back to the conditioning approach:

$$y^{\star} = \mathrm{E}(y|\boldsymbol{x}) = \beta_0 + \boldsymbol{\beta}^T\boldsymbol{x}$$

$$\text{Var}(\varepsilon) = \text{Var}(y|\boldsymbol{x}) = \sigma_y^2(1-\Omega^2)$$

We know that the mean squared prediction error for $y^\star$ is $\text{E}((y-y^\star)^2) = \text{Var}(\varepsilon)$ and that this is the minimal irreducible error. Hence, we may use $\text{Var}(\varepsilon)$ as the *minimum* variability for the prediction.

The corresponding prediction interval is

$$\left[y^\star(\boldsymbol{x}^{\text{test}}) \pm c \times \text{SD}(\varepsilon)\right]$$

where c is some suitable constant (e.g. 1.96 for symmetric 95% normal intervals).

However, please note that the prediction interval constructed in this fashion will be an *underestimate*. The reason is that this assumes that we employ $y^\star = \beta_0 + \boldsymbol{\beta}^T\boldsymbol{x}$ but in reality we actually use $\hat{y} = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}^T\boldsymbol{x}$ for prediction — note the estimated coefficients! We recall from an earlier chapter (best linear predictor) that this leads to increase of MSPE compared with using the optimal β_0 and $\boldsymbol{\beta}$.

Thus, for better prediction intervals we would need to consider the mean squared prediction error of $\hat{y}$ that can be written as $\text{E}((y-\hat{y})^2) = \text{Var}(\varepsilon) + \delta$ where δ is an **additional error term due to using an estimated rather than the true regression function**. δ typically declines with $1/n$ but can be substantial for small n (in particular as it usually depends on the number of predictors d).

For more details on this we refer to later modules on regression.

19.2 Variable importance and prediction

Another key question in regression modeling is to find out predictor variables $x_1, x_2, \ldots, x_d$ are actually important for predicting the outcome y.

$\rightarrow$ We need to study variable importance measures (VIM).

19.2.1 How to quantify variable importance?

A variable x_i is **important** if it **improves prediction** of the response y.

Recall variance decomposition:

$$\text{Var}(y) = \sigma_y^2 = \underbrace{\sigma_y^2\Omega^2}_{\text{explained variance}} + \underbrace{\sigma_y^2(1-\Omega^2)}_{\text{unexplained/residual variance} = \text{Var}(\varepsilon)}$$

- Ω^2 squared multiple correlation $\in [0,1]$
- Ω^2 large $\rightarrow$ 1 predictor variables explain most of σ_y^2
- Ω^2 small $\rightarrow$ 0 linear model fails and predictors do not explain variability
- $\Rightarrow$ If a predictor helps to $\begin{matrix}\text{increase explained variance}\\ \text{decrease unexplained variance}\end{matrix}$ then it is important!
- $\Omega^2 = \boldsymbol{P}_{yx}\boldsymbol{P}_{xx}^{-1}\boldsymbol{P}_{xy} \hat{=}$ a function of the X!

VIM: which predictors contribute most to Ω^2

19.2.2 Some candidates for VIMs

1. The regression coefficients $\boldsymbol{\beta}$
   - $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\boldsymbol{xx}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}y} = \boldsymbol{V}_{\boldsymbol{x}}^{-1/2}\boldsymbol{P}_{\boldsymbol{xx}}^{-1}\boldsymbol{P}_{\boldsymbol{x}y}\sigma_y$
   - Not a good VIM since $\boldsymbol{\beta}$ contains the scale!
   - Large $\hat{\beta}_i$ does not indicate that x_i is important.
   - Small $\hat{\beta}_i$ does not indicate that x_i is not important.
2. Standardised regression coefficients $\boldsymbol{\beta}_{\text{std}}$
   - $\boldsymbol{\beta}_{\text{std}} = \boldsymbol{P}_{\boldsymbol{xx}}^{-1}\boldsymbol{P}_{\boldsymbol{x}y}$
   - implies $\text{Var}(y) = 1, \text{Var}(x_i) = 1$
   - These do not contain the scale (so better than $\hat{\beta}$)
   - But still unclear how this relates to decomposition of variance
3. Squared marginal correlations ρ^2_{y,x_i}

   Consider case of uncorrelated predictors: $\boldsymbol{P}_{\boldsymbol{xx}} = \boldsymbol{I}$ (no correlation among x_i)

$$\Rightarrow \Omega^2 = \boldsymbol{P}_{\boldsymbol{y}y}\boldsymbol{P}_{\boldsymbol{y}y} = \sum_{i=1}^{d} \rho^2_{y,x_i}$$

   $\rho^2_{y,x_i} = \text{Cor}(y, x_i)$ is the marginal correlation between y and x_i, and Ω^2 is (for uncorrelated predictors) the sum of squared marginal correlations.
   - If $\boldsymbol{P}_{\boldsymbol{xx}} = \boldsymbol{I}$, then *ranking* predictors by ρ^2_{y,x_i} is optimal!
   - The predictor with largest marginal correlation reduces the unexplained variance most!
   - good news: even if there is weak correlation among predictors the marginal correlations are still good as VIM (but then they will not perfectly add up to Ω^2)
   - Advantage: very simple but often also very effective.
   - Caution! If there is strong correlation in $\boldsymbol{P}_{\boldsymbol{xx}}$, then there is **colinearity** (in this case it is oftern best to remove one of the strongly correlated variables, or to merge the correlated variables).

Often, ranking predictors by their squared marginal correlations is done as a prefiltering step (independence screening).

19.3 Regression t-scores.

So far, we discussed three obvious candidates for for variable importance measures (regression coefficients, standardised regression coefficients, marginal correlations). However, of these only marginal correlation between the response y and each predictor x_i are useful for variable ranking (due to the direct link of multiple correlation and explained variation).

In this section we consider a further quantity, the **regression** $t-$**score**:

Recall that ML estimation of the regression coefficients yields

- a point estimate $\hat{\boldsymbol{\beta}}$
- the (asymptotic) variance $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})$
- the asymptotic normal distribution $\hat{\boldsymbol{\beta}} \overset{a}{\sim} N_d(\boldsymbol{\beta}, \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}))$

Corresponding to each predictor x_i we can construct from the above a t-score

$$t_i = \frac{\hat{\beta}_i}{\widehat{\mathrm{SD}}(\hat{\beta}_i)}$$

where the standard deviations are computed by $\widehat{\mathrm{SD}}(\hat{\beta}_i) = \mathrm{Diag}(\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}))_i$. Asymptotically, under the null hypthesis that $\beta_i = 0$ we have

$$t_i \overset{a}{\sim} N(0,1)$$

so that corresponding (symmetric) p-values can be computed by $p = 2\Phi(-|t_i|)$. For finite n and normal assumption one can show that t_i actually follows a t-distribution:

$$t_i \sim t_{n-d-1}$$

Regression t-scores can thus be used to test whether a regression coefficient is zero, and also to rank predictors either by $|t_i|$ or the p-values (both will lead to the same ranking of course).

Note that by means of construction the regression t-scores do not depend on the data variance, so when variables are rescaled this will not affect the corresponding regression t-score.

Furthermore, if $\widehat{\mathrm{SD}}(\hat{\beta}_i)$ is small, then the regression t-score t_i can be large even if $\hat{\beta}_i$ is small!

19.3.1 Computation

When you perform regression analysis in R (or any other software), you will get to see the following table:

$\hat{\beta}_i$ Estimated repression coefficient	$\widehat{\mathrm{SD}}(\hat{\beta}_i)$ Error of $\hat{\beta}_i$	$t_i = \frac{\hat{\beta}_i}{\widehat{\mathrm{SD}}(\hat{\beta}_i)}$ t-score computed from first two columns	p-values for t_i based on t-distribution	Indicator of Significance * 0.9 ** 0.95 *** 0.99

Thus, using regression t-scores is very common, because they will be computed automatically by most software!

19.3.2 Connection with partial correlation

The deeper reason why ranking predictors by regression t-scores and associated p-values is useful is their link with **partial correlation**.

In particular, the regression t-score can be 1:1 transformed into the (estimated) partial correlation

$$\hat{\rho}_{y,x_i|x_{j\neq i}} = \frac{t_i}{\sqrt{t_i^2 + df}}$$

with $df = n - d - 1$, and it can be shown that the p-values for testing that $\beta_i = 0$ are exactly the same as the p-values for testing that the partial correlation $\rho_{y,x_i|x_{j\neq i}}$ vanishes!

Therefore, ranking the predictors x_i by regression t-scores leads to exactly the same ranking and p-values as partial correlation!

19.4 Further approaches for variable selection

In addition to ranking by marginal and partial correlation, there are many other approaches for variable selection in regression!

a) Search-based methods:

- search through subsets of linear models for d variables, ranging from full model (including all predictors) to the empty model (includes no predictor) and everything inbetween.
- Problem: exhaustive search not possible even for relatively small d as space of models is very large!
- Therefore heuristic approaches such as forward selection (adding predictors), backward selection (removing predictors), or monte-carlo random search are employed.
- Problem: maximum likelihood cannot be used for choosing among the models - since ML will always pick the best model. Therefore, penalised ML criteria such as AIC or Bayesian criteria are often employed instead.

b) Integrative estimation and variable selection:

- there are methods that fit the regression model and perform variable selection *simultaneously*.
- the most well-known approach of this type is "lasso" regression (Tibshirani 1996)
- This applies a (generalised) linear model with ML plus L1 penalty.
- Alternative: Bayesian variable selection and estimation procedures

c) Entropy-based variable selection:

As seen above, two of the most popular approaches in linear models are based on correlation, eithermarginal correlation or partial correlation (via regression t-scores).

Correlation measures can be generalised to non-linear settings. One very popular measure is the **mutual information** which is computed using the KL divergence. In case of two variables x and y with joint normal distribution and correlation ρ the mutual information is a function of the

correlation:

$$\mathrm{MI}(x, y) = \frac{1}{2} \log(1 - \rho^2)$$

In regression he mutual information between the response y and predictor x_i is $\mathrm{MI}(y, x_i)$, and this widely used for feature selection, in particular in machine learning.

Appendix

Appendix A

Refresher

Statistics is a mathematical science that requires practical use of tools from probability, vector and matrices, analysis etc.

Here we briefly list some essentials that are needed for "Statistical Methods". Please familiarise yourself (again) with these topics.

A.1 Vectors and matrices

Vector and matrix notation.

Vector algebra.

Eigenvectors and eigenvalues for a real symmetric matrix.

Eigenvalue (spectral) decomposition of a real symmetric matrix.

Positive and negative definiteness of a real symmetric matrix (containing only positive or only negative eigenvalues).

Singularity of a real symmetric matrix (containing one or more eigenvalues identical to zero).

Singular value decomposition of a real matrix.

A.2 Functions

A.2.1 Gradient

The **nabla operator** (also known as **del operator**) is the *row* vector

$$\nabla = (\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_d}) = \frac{\partial}{\partial \boldsymbol{x}}$$

containing the first order partial derivative operators.

The **gradient** of a scalar-valued function $f(\boldsymbol{x})$ with vector argument $\boldsymbol{x} = (x_1, \ldots, x_d)^T$ is also a *row* vector (with d columns) and can be expressed

using the nabla operator

$$\nabla f(\boldsymbol{x}) = \left(\frac{\partial f(\boldsymbol{x})}{\partial x_1}, \ldots, \frac{\partial f(\boldsymbol{x})}{\partial x_d} \right) = \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \text{grad} f(\boldsymbol{x}) \,.$$

Note the various notations for the gradient.

Examples:

- $f(\boldsymbol{x}) = \boldsymbol{a}^T \boldsymbol{x} + b$. Then $\nabla f(\boldsymbol{x}) = \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \boldsymbol{a}^T$.
- $f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{x}$. Then $\nabla f(\boldsymbol{x}) = \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = 2\boldsymbol{x}^T$.
- $f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}$. Then $\nabla f(\boldsymbol{x}) = \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \boldsymbol{x}^T (\boldsymbol{A} + \boldsymbol{A}^T)$.

A.2.2 Hessian matrix

The matrix of all second order partial derivates of scalar-valued function with vector-valued argument is called the **Hessian matrix** and is computed by double application of the nabla operator:

$$\nabla^T \nabla f(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1^2} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_d \partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_d^2} \end{pmatrix} = \left(\frac{\partial f(\boldsymbol{x})}{\partial x_i \partial x_j} \right) = \left(\frac{\partial}{\partial \boldsymbol{x}} \right)^T \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} \,.$$

By construction it is square and symmetric.

A.2.3 Conditions for local maximum of a function

Function has one variable:

i) First derivative is zero at maximum.
ii) Second derivative is negative at maximum (negative curvature).

Function has several variables:

i) Gradient vanishes at maximum.
ii) Negative definite Hessian matrix at maximum (all eigenvalues of Hessian matrix are negative).

A.2.4 Linear and quadratic approximation

Taylor series of first / second order.

Applied to scalar-valued function of a scalar:

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2$$

With $x = x_0 + \varepsilon$ this can be written as

$$f(x_0 + \varepsilon) \approx f(x_0) + f'(x_0)\, \varepsilon + \frac{1}{2} f''(x_0)\, \varepsilon^2$$

Applied to scalar-valued function of a vector:

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^T \nabla^T \nabla f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0)$$

With $\boldsymbol{x} = \boldsymbol{x}_0 + \boldsymbol{\varepsilon}$ this can be written as

$$f(\boldsymbol{x}_0 + \boldsymbol{\varepsilon}) \approx f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)\boldsymbol{\varepsilon} + \frac{1}{2}\boldsymbol{\varepsilon}^T \nabla^T \nabla f(\boldsymbol{x}_0)\boldsymbol{\varepsilon}$$

A.2.5 Functions of matrices

Matrix inverse, matrix square root etc. of symmetric real matrices.

Computation via eigenvalue decomposition i.e. apply function such as inverse, sqrt etc. on the eigenvalues.

In this course we do not actually compute matrix functions, but we will use matrix notation for matrix square roots, so you do need to know that it is exists and that it is not the same as taking the square root of the matrix entries.

Trace and determinant of a square matrix.

Connection with eigenvalues (trace = sum of eigenvalues, determinant = product of eigenvalues).

A.3 Probability

A.3.1 Law of large numbers:

- By the strong law of large numbers the empirical distribution $\hat{F}_n$ converges to the true underlying distribution F as $n \to \infty$ almost surely:

  $$\hat{F}_n \xrightarrow{a.s.} F$$

  The Glivenko–Cantelli theorem asserts that the convergence is uniform. Since the strong law implies the weak law we also have convergence in probability:

  $$\hat{F}_n \xrightarrow{P} F$$

- Correspondingly, for $n \to \infty$ the average $\mathrm{E}_{\hat{F}_n}(h(X)) = \frac{1}{n}\sum_{i=1}^{n} h(x_i)$ converges to the expectation $\mathrm{E}_F(h(X))$.

A.3.2 Jensen's inequality

$$\mathrm{E}(h(X)) \geq h(\mathrm{E}(X))$$

for a *convex* function $h(x)$.

A function is convex if $h''(x) \geq 0$. Note: if $h(x)$ is convex, then $-h(x)$ is *concave*.

A.3.3 Transformation of univariate densities

For a general coordinate transformation $y = h(x)$ the backtransformation is $x = h^{-1}(y)$.

The transformation of the infinitesimal volume element is $dy = |\frac{dy}{dx}|dx$.

The transformation of the density is $f_y(y) = \left|\frac{dy}{dx}\right|^{-1} f_x(h^{-1}(y))$.

A.3.4 Normal distribution

Univariate normal distribution:

$x \sim N(\mu, \sigma^2)$ with $\mathrm{E}(x) = \mu$ and $\mathrm{Var}(x) = \sigma^2$.

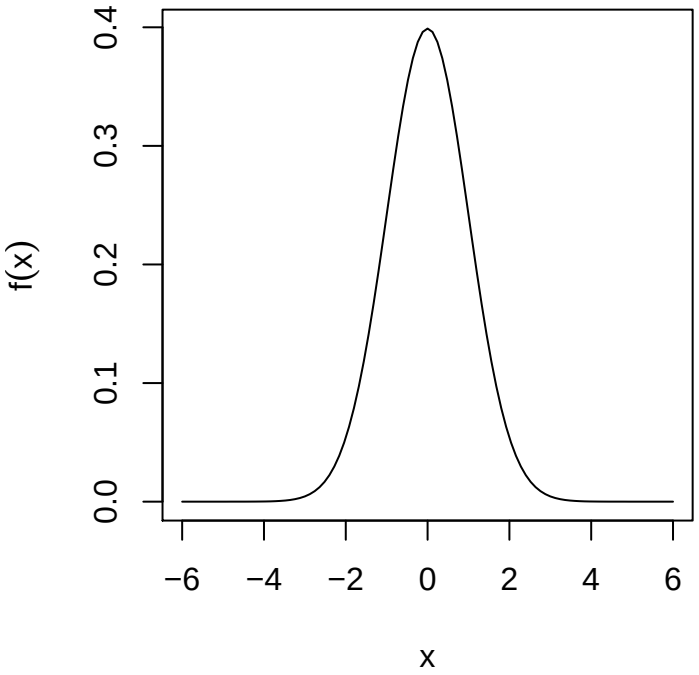
Probability density function (PDF):

$$f(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

In R the density function is called `dnorm()`.

The standard normal distribution is $N(0, 1)$ with mean 1 and variance 1.

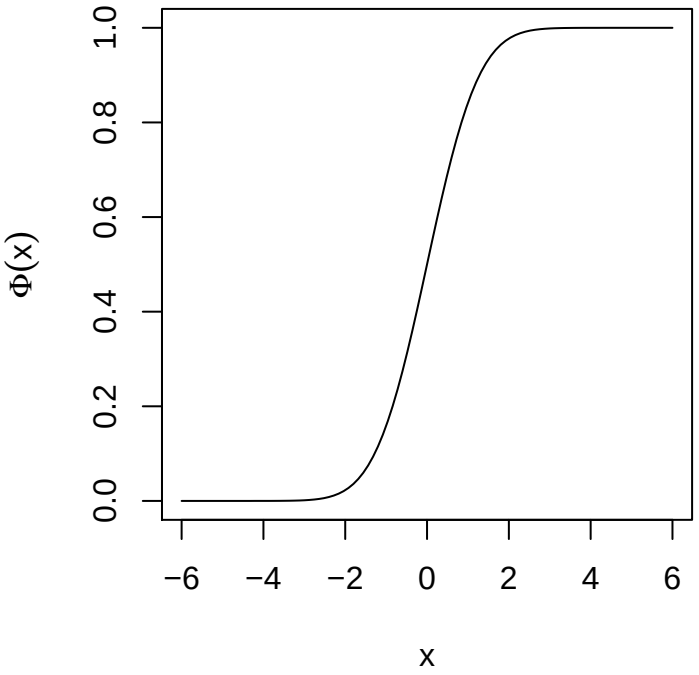
Plot of the PDF of the standard normal:



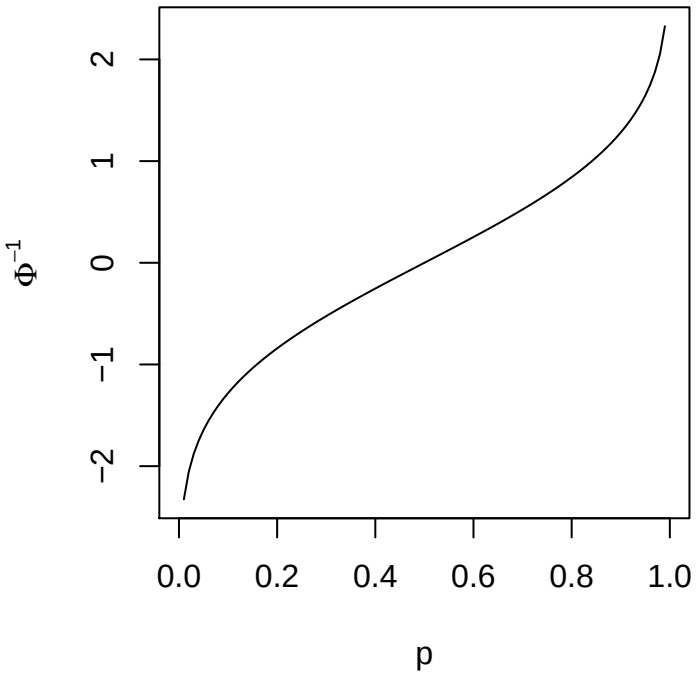
The cumulative distribution function (CDF) of the standard normal $N(0, 1)$ is

$$\Phi(x) = \int_{-\infty}^{x} f(x'|\mu = 0, \sigma^2 = 1)dx'$$

There is no analytic expression for $\Phi(x)$. In R the function is called `pnorm()`.



The inverse $\Phi^{-1}(p)$ is called the quantile function of the standard normal. In R the function is called `qnorm()`.



A.3.5 Chi-squared distribution

Asume m independent standard normal random variables

$$z_1, z_2, \ldots, z_m \sim N(0,1)$$

Then the sum of the squares

$$x = \sum_{i=1}^{m} z_i^2$$

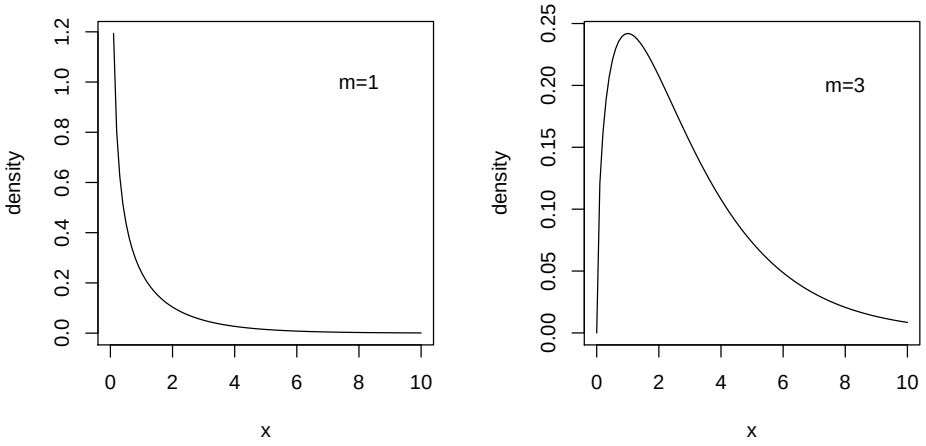
is a chi-squared random variable $x \sim \chi^2_m$ with degree of freedom m and $x \geq 0$.

The mean of a χ^2_m distributed random variable x is $\mathrm{E}(x) = m$ and the variance $\mathrm{Var}(x) = 2m$.

The chi-squared distribution is a special case (assuming $\sigma^2 = 1$) of the scaled chi-squared distribution $\sigma^2\chi^2_m$ that arises if the $z_i \sim N(0, \sigma^2)$ have variance σ^2. The mean and variance of a scaled chi-squared distributed variable is $\mathrm{E}(x) = m\sigma^2$ and $\mathrm{Var}(x) = 2m\sigma^4$.

The Gamma distribution $Gamma(\alpha, \beta)$ is another name of the scaled chi-squared distribution but by convention is uses a different parameterisation with shape parameter α and scale parameter β. The scaled chi-squared distribution $\sigma^2\chi^2_m$ equals $Gamma(\frac{m}{2}, 2\sigma^2)$ and the chi-squared distribution χ^2_m equals $Gamma(\frac{m}{2}, 2)$.

Density of the chi-squared distribution for degrees of freedom $m = 1$ and $m = 3$:



In R the density of the chi-squared distribution is given by `dchisq()`. The cumulative density function is `pchisq()` and the quantile function is `qchisq()`.

The density of the Gamma distribution (aka scaled chi-squared distribution) is available in by `dgamma()`. The cumulative density function is `pgamma()` and the quantile function is `qgamma()`.

A.4 Statistics

A.4.1 Statistical learning

The aim in statistics - data science - statistics - machine learning is to learn from data (from experiments, observations, measurements) to learn about and understand the world.

Specifically, to identify the best model(s) for the data in order to

- to explain the current data, and
- to enable good prediction of future data

Note that it is easy to get models that only explain the data but do not predict well!

This is called *overfitting* the data and happens in particular if the model is overparameterized for the amount of data available.

Specifically, we have data $x_1, \ldots, x_n$ and models $f(x|\theta)$ that are indexed the parameter θ.

Often (but not always) θ can be interpreted and/or is associated with some property of the model.

If there is only a single parameter we write θ (scalar parameter). For a parameter vector we write $\boldsymbol{\theta}$ (in bold type).

A.4.2 Point and interval estimation

- There is a parameter θ of interest in a model
- we are uncertain about this parameter (i.e. we don't know the exact value)
- we would like to learn about this parameter by observing data $x_1, \ldots, x_n$ from the model

Estimation:

- An **estimator for** θ is a function $\hat{\theta}(x_1, \ldots, x_n)$ that maps the data (input) to a "guess" (output) about θ.

- A **point estimator** provides a single number for each parameter
- An **interval estimator** provides a set of possible values for each parameter.

A.4.3 Sampling properties of a point estimator $\hat{\theta}$

A point estimator $\hat{\theta}$ depends on the data, hence it has sampling variation (i.e. estimate will be different for a new set of observations)

Thus $\hat{\theta}$ can be seen as a random variable, and its distribution is called sampling distribution (accross different experiments).

Properties of this distribution can be used to evaluate how far the estimator deviates (on average across different experiments) from the true value:

$$\begin{array}{rrl} \text{Bias:} & \text{Bias}(\hat{\theta}) & = \mathrm{E}(\hat{\theta}) - \theta \\ \text{Variance:} & \text{Var}(\hat{\theta}) & = \mathrm{E}\left((\hat{\theta} - \mathrm{E}(\hat{\theta}))^2\right) \\ \text{Mean squared error:} & \text{MSE}(\hat{\theta}) & = \mathrm{E}((\hat{\theta} - \theta)^2) \\ & & = \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2 \end{array}$$

The last identity about MSE follows from $\mathrm{E}(X^2) = \text{Var}(X) + \mathrm{E}(X)^2$.

At first sight it seems desirable to focus on unbiased (for finite n) estimators. However, requiring strict unbiasedness is not always a good idea!

In many situations it is better to allow for some small bias and in order to achieve a smaller variance and an overall total smaller MSE. This is called *bias-variance tradeoff* — as more bias is traded for smaller variance (or, conversely, less bias is traded for higher variance)

A.4.4 Asymptotics

Typically, Bias, Var and MSE all decrease with increasing sample size so that with more data $n \to \infty$ the errors become smaller and smaller.

The typical rate of decrease of variance of a good estimator is $\frac{1}{n}$. Thus, when sample size is doubled the variance is divided by 2 (and the standard deviation is divided by $\sqrt{2}$).

Consistency: $\hat{\theta}$ is called consistent if

$$\text{MSE}(\hat{\theta}) \longrightarrow 0 \text{ with } n \to \infty$$

.

Consistency implies we recover the true model in the limit of infinite data and if the model class contains the true model.

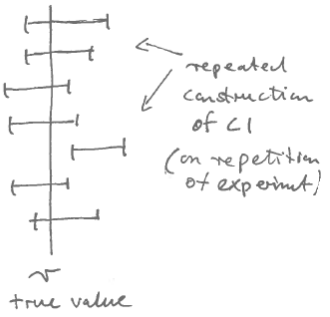
Consistency is a *minimum* essential requirement for any reasonable estimator! Of all consistent estimators we typically prefer the estimator that is most efficient (i.e. with fasted decrease in MSE) and that thus has smallest variance and/or MSE for given finite n.

Note that if the model class does not contain the true model then strict consistency cannot be achived but we still wish to get at least as close as possible to the true model.

A.4.5 Confidence intervals

- A **confidence** interval (CI) is an **interval estimate** with a **frequentist** interpretation.
- Definition of **coverage** κ of a CI: how often (in repeated identical experiment) does the estimated CI overlap the true parameter value θ
  - Eg.: Coverage $\kappa = 0.95$ (95%) means that in 95 out of 100 case the estimated CI will contain the (unknown) true value (i.e. it will "cover" θ).

Illustration of the repeated construction of a CI for θ:



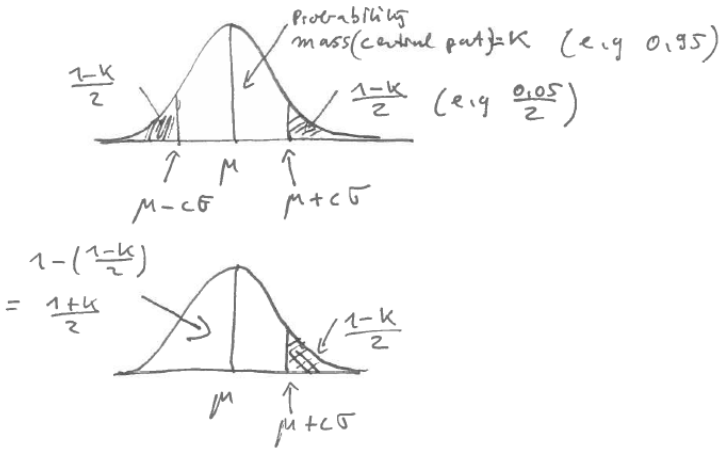
- Note that a CI is actually an **estimate**: $\widehat{\text{CI}}(x_1, \ldots, x_n)$, i.e. it depends on data and has a random (sampling) variation.

- A good CI has high coverage and is compact.

Note: the coverage probability is **not** the probability that the true value is contained in a given estimated interval (that would be the Bayesian *Credible* Interval).

A.4.6 Symmetric normal confidence interval

For a normally distributed univariate random variable it is straightforward to construct a symmetric two-sided CI with a given desired coverage κ.



For a normal random variable $X \sim N(\mu, \sigma^2)$ with mean μ and variance σ^2 and density function $f(x)$ we can compute the probability

$$\Pr(X \leq \mu + c\sigma) = \int_{-\infty}^{\mu+c\sigma} f(x)dx = \Phi(c) = \frac{1+\kappa}{2}$$

Note $\Phi(c)$ is the cumulative distribution function (CDF) of the standard normal $N(0, 1)$:

From the above we obtain the critical point c from the quantile function, i.e. by inversion of Φ:

$$c = \Phi^{-1}\left(\frac{1+\kappa}{2}\right)$$

The following table lists c for the three most commonly used values of κ - it is useful to memorise these values!

Coverage κ	Critical value c
0.9	1.64
0.95	1.96
0.99	2.58

A **symmetric standard normal CI** with nominal coverage κ for

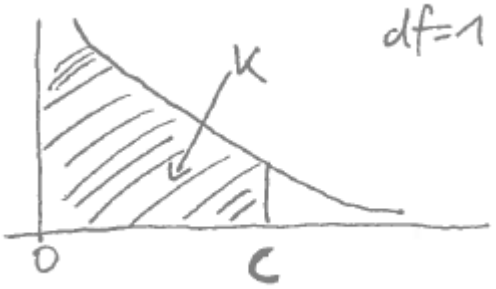
- a scalar parameter θ
- with normally distributed estimate $\hat{\theta}$ and
- with estimated standard deviation $\widehat{\mathrm{SD}}(\hat{\theta}) = \hat{\sigma}$

is then given by

$$\widehat{\mathrm{CI}} = [\hat{\theta} \pm c\hat{\sigma}]$$

where c is chosen for desired coverage level κ.

A.4.7 Confidence interval for chi-squared distribution



As for the normal CI we can compute critical values but for the chi-squared distribution we use a one-sided interval:

$$\Pr(X \leq c) = \kappa$$

As before we get c by the quantile function, i.e. by inverting the CDF of the chi-squared distribution.

The following list the critical values for the three most common choice of κ for $m = 1$ (one degree of freedom):

Coverage κ	Critical value c ($m = 1$)
0.9	2.71
0.95	3.84
0.99	6.63

A one-sided CI with nominal coverage κ is then given by $[0, c]$.

Appendix B

Further study

In this module we can only touch the surface of likelihood and Bayes inference. As a starting point for further reading the following text books are recommended.

B.1 Recommended reading

- Held and Bové (2014) *Applied Statistical Inference: Likelihood and Bayes*. Springer.
- Faraway (2015) *Linear Models with R (second edition)*. Chapman and Hall/CRC.

B.2 Additional references

- Wood (2015) *Core Statistics*. Cambridge University Press.
- Gelman et al. (2014) *Bayesian data analysis (3rd edition)*. CRC Press.

Bibliography

Domingos, P. 2015. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. Basic Books.

Faraway, J. J. 2015. *Linear Models with R*. 2nd ed. Chapman; Hall/CRC.

Gelman, A., J. B. Carlin, H. A. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2014. *Bayesian Data Analysis*. 3rd ed. CRC Press.

Held, L., and D. S. Bové. 2014. *Applied Statistical Inference: Likelihood and Bayes*. Springer.

Wood, S. 2015. *Core Statistics*. Cambridge University Press.