STATISTICAL AND MATHEMATICAL FOUNDATIONS

MODULE LEARNING OBJECTIVES

In this module, we provide, in very broad terms, some fundamental mathematical and statistical background required for data analysis and model building with practical applications.

Participants will become acquainted will **key concepts**, to allow for future learning.

This introduction is not meant to replace formal training and is at best **incomplete**; please consult the references for further details.

OUTLINE

1. Modeling
2. Distributions
3. Central Limit Theorem
4. Estimation
5. Bayes’ Theorem
6. Matrix Algebra
7. Eigenvalues and Eigenvectors
8. Regression
9. Optimization

MODELING

STATISTICAL AND MATHEMATICAL FOUNDATIONS

LEARNING OBJECTIVES

Understand the difference between modeling from first principles and statistical modeling.

Working knowledge of the modeling process.

Increase awareness of modeling pitfalls and challenges.

[http://dailycatdrawings.tumblr.com; #309]

Real World

Model

Theory

Identification of details relevant to **description** and **translation** of real-world objects into model variables

MODELS IN GENERAL

First principles modeling

- examine a system
- write down a set of rules/equations that describe the essence of the system
- ignore complicating details that are “less” important

Statistical modeling

- typically a set of equations with parameters
- parameters are learned (model is “trained”) using multiple data observations
- data sample vs. population

MODELING HEURISTICS

In a sense, modeling is a **straightforward** (and **formulaic**?) process, guided by **intuition** and **experience** at each step.

Basic steps in building a statistical model:

- **defining the goals**
  - what are we trying to achieve?
  - under what situations will the model be used and what is the outcome we are trying to predict?
- **gathering data**
  - what data is available?
  - how many records will we have?
  - generally, modelers want as much data as possible

MODELING HEURISTICS

Basic steps in building a statistical model: (continued)

- **deciding on the model structure**

  - should we run a linear regression, logistic regression, or a nonlinear model? Which kind?

  - choices of model structure require experience and deep knowledge of the strength and weaknesses of each technique

- **preparing the data**

  - assemble data into appropriate form for the model

  - encode the data into inputs, using expert knowledge as much as possible

  - separate the data into the desired training, testing, and validation sets

MODELING HEURISTICS

Basic steps in building a statistical model: (continued)

- **selecting and removing features**

  - variables are examined for model importance and selected or eliminated

  - a list of candidate appropriate variables are ordered by importance

- **building candidate models**

  - begin with baseline linear models and try to improve using more complex nonlinear models

  - keep in mind the environment in which the model will be implemented

- **finalizing the model**

  - select among the candidates the most appropriate model to be implemented

- **implementing and monitoring**

  - embed the model into necessary system process; implement monitoring steps to examine the model performance

MODELING PITFALLS

Common pitfalls surrounding the modeling process:

- **defining the goals**

  - lack of clarity around problem definition

  - lack of understanding of how and where the model will be used

- **gathering data**

  - using data that is too old or otherwise not relevant going forward

  - not considering additional key data sources or data sets that might be available

- **deciding on the model structure**

  - using a modeling methodology that is not appropriate for the nature of the data (sizes, dimensions, noise...)

MODELING PITFALLS

Common pitfalls surrounding the modeling process: (continued)

- **preparing the data**

  - not cleaning or considering outliers

  - not properly scaling data

  - not giving enough thought to building special expert variables

  - not having data from important categories of records

- **selecting and eliminating features**

  - keeping too many variables, making it hard for modeling, interpretation, implementation, or model maintenance

  - too much reliance on simply eliminating correlated variables

MODELING PITFALLS

Common pitfalls surrounding the modeling process: (continued)

- **building candidate models**
  - overfitting
  - not doing proper training/testing as one examines candidate models
  - not doing a simpler linear regression to use as baseline
- **finalizing the model**
  - not rebuilding the final model optimally using all the appropriate data
  - improperly selecting the final model without consideration to some implementation constraints
- **implementing and monitoring**
  - errors in implementation process: data input streams, variable encodings, algorithm mistakes
  - not monitoring model performance

DISTRIBUTIONS

STATISTICAL AND MATHEMATICAL FOUNDATIONS

LEARNING OBJECTIVES

What questions can you use to help you pick a model distribution for a data feature?

What are some commonly encountered pdfs?

What are the mean and variance of some common pdfs?

When do we need to use joint distributions?

DATA AND DISTRIBUTIONS

If a data feature can be characterized by a distribution, consider asking **four basic questions**:

1. Can the variable only take on **discrete** values? **continuous** values?
   - whether a taxpayer's file is audited or not is a *discrete* variable but the corrected amount from the audit is a *continuous* variable
2. Is the data distribution **symmetric**?
   - If not, in which **direction** does the asymmetry lie?
   - Are **right-** and **left-outliers** equally likely?

DATA AND DISTRIBUTIONS

3. Does the variable have theoretical **upper** and **lower limits**?
   - Some items like age or height cannot be smaller than zero
   - Some items like operating margins cannot exceed a value (100% in this case)
4. How likely is it to observe **extreme values** in the distribution?
   - in some data, extreme values occur infrequently whereas in others, they occur more often

How would these questions have to change when dealing with **joint distributions**?

FUNDAMENTAL DISTRIBUTIONS

Empirical distributions are often approximated by **parametric distributions**, defined *via* a **probability density function** (pdf) and a set of parameters that must be learned from the data.

The basic distributions of data analysis are:

- the **uniform** distribution $U(a,b)$ on the interval $[a,b]$ or $U(x_1,\ldots,x_n)$ on the discrete set $\{x_1,\ldots,x_n\}$, potentially the simplest
- the **normal** distribution $N(\mu,\sigma^2)$ on the real line $\mathbb{R}$, possibly the most frequently used (not always aptly so)
- a wide variety of **special** distributions that are used in applications ranging from consumer modeling and finance to operation research (**Poisson**, **exponential**, **log-normal**, **binomial**, etc.)

EXPECTATION AND MOMENTS

Given a pdf f and a function $g(X)$, the **expectation** $\mathrm{E}_f(g(X))$ **of** g **under** f is the **weighted average**

$$\mathrm{E}_f\big(g(X)\big) = \int_\Omega g(X)f(X)\,dX, \text{ where } \Omega = \mathrm{dom}(f).$$

The **moments** of a distribution are defined as

$$m_i = \mathrm{E}\big(X^i\big), \text{ for } i = 0, \ldots,$$

Note that $m_0 = 1$, by definition. The **mean** and **variance** of the distribution are given by $m_1 = \mathrm{E}(X)$ and $m_2 - m_1^2 = \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2$, respectively.

distribution	pdf $f(x)$	mean	variance	notes
uniform $U(a,b)$	$\frac{1}{b-a}$ for $a \le x \le b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	most languages provide rand # generators for $U(a,b)$; used to generate r.v. with other distributions
Gaussian $N(\mu,\sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ for $x \in \mathbb{R}$	μ	σ^2	if $X \sim N(\mu,\sigma^2)$, then $\frac{X-\mu}{\sigma} \sim N(0,1)$ (and *vice-versa*); very commonly used
Poisson $P(\lambda), \lambda \ge 0$	$\frac{\lambda^x}{x!} e^{-\lambda}$ for $x = 0,1,2,\ldots$	λ	λ	estimates the # of events that occur in a continuous time interval (# of calls received in 1-hour intervals)
binomial $\mathcal{B}(N,p), N \in \mathbb{N}$, $p \in [0,1]$	$\binom{N}{x} p^x (1-p)^{N-x}$ for $x = 0,1,\ldots,N$	Np	$Np(1-p)$	describes the probability of exactly x successes in N independent trials if the probability of a success in a single trial is p (# of heads in N coin tosses)
log-normal $\Lambda(\mu,\sigma^2)$	$\frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln x-\mu}{\sigma}\right)^2}$ for $x > 0$	$e^{(\mu+\sigma^2/2)}$	$e^{(2\mu+\sigma^2)}\left[e^{\sigma^2} - 1\right]$	if $\ln X \sim N(\mu,\sigma^2)$, then $X \sim \Lambda(\mu,\sigma^2)$ (and *vice-versa*); positively skewed

JOINT DISTRIBUTIONS

Univariate distributions are useful modeling tools, especially when the variables under consideration are **independent**.

In practice, that is not usually the case. A **joint distribution** $P(X_1,\ldots,X_n)$ gives the probability that each of $X_1,\ldots,X_n$ falls in a given range. The **multivariate normal distribution** $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ has pdf

$$f(x_1,\ldots,x_n) := f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ the covariance matrix.

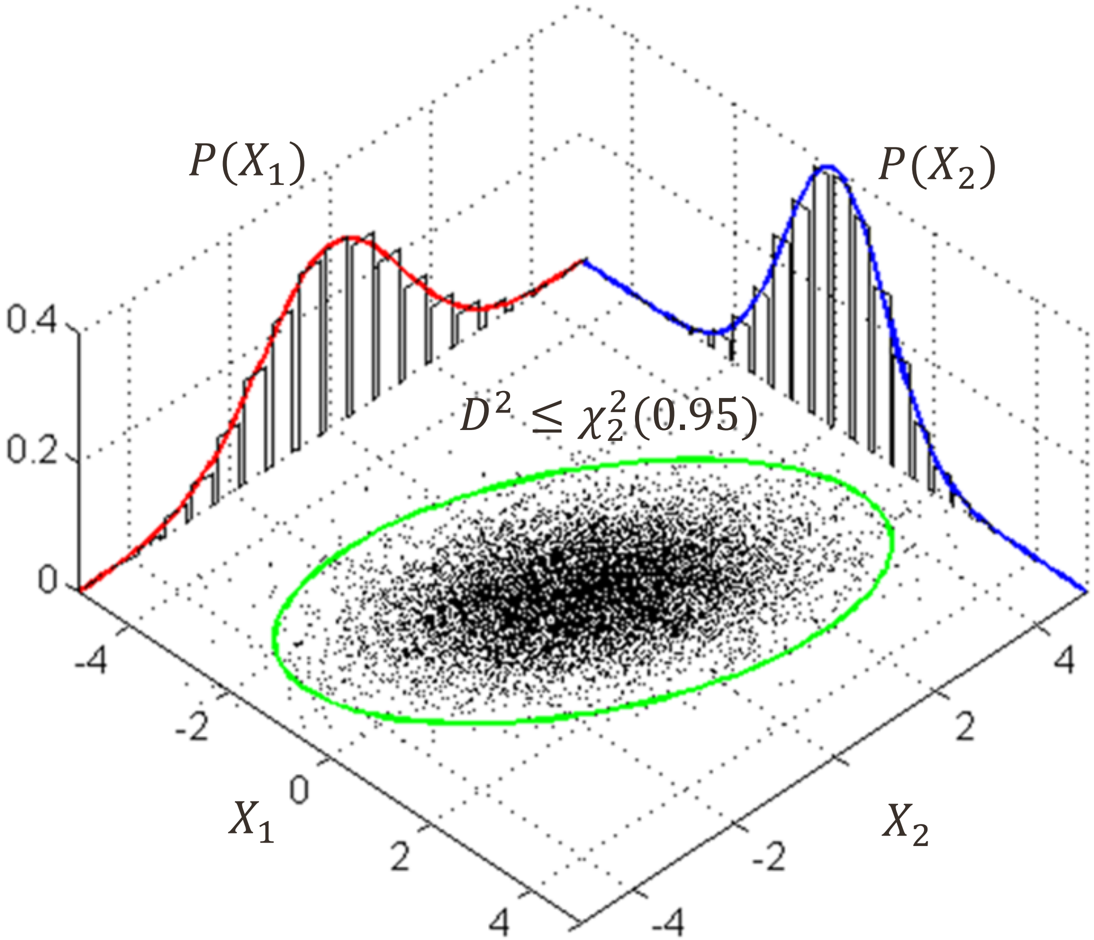
[image taken from Wikipedia]

JOINT DISTRIBUTIONS

If $\mathbf{\Sigma}$ is positive definite, the multivariate normal is **non-degenerate**.

$D = \sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$ is the **Mahalanobis** distance.

To generate a sample $\boldsymbol{x}$ from $N(\boldsymbol{\mu},\mathbf{\Sigma})$, let $\boldsymbol{z}\sim N(\mathbf{0},\mathbf{I})$ and set $\boldsymbol{x}=\boldsymbol{\mu}+\boldsymbol{A}\boldsymbol{z}$, where $\boldsymbol{A}\boldsymbol{A}^T=\mathbf{\Sigma}$ is the Cholesky decomposition.



EXERCISES

Write R and/or Python code that lets you draw “random” samples from the various distributions discussed in this section.

CENTRAL LIMIT THEOREM

STATISTICAL AND MATHEMATICAL FOUNDATIONS

LEARNING OBJECTIVES

What is the central limit theorem?

When is the central limit theorem relevant?

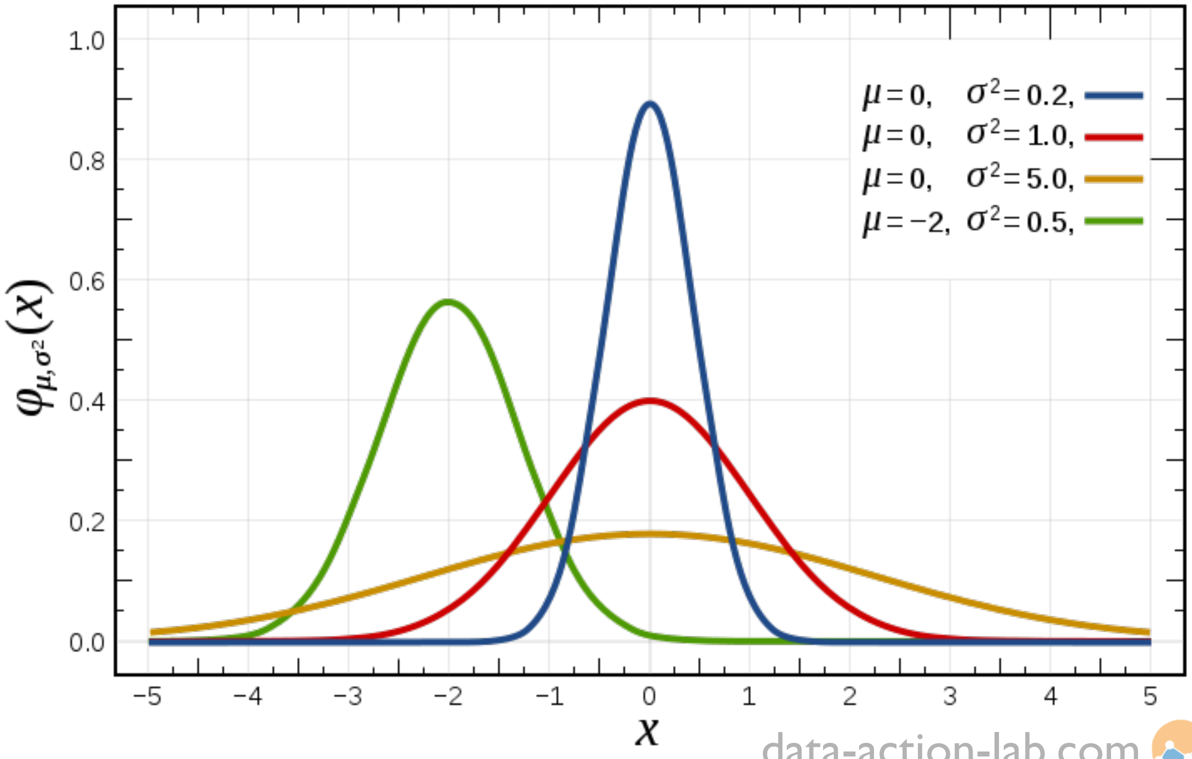
[image taken from Wikipedia]

NORMAL DISTRIBUTION

$N(\mu, \sigma^2)$ is **fully characterized** by the mean μ and the standard deviation σ, which reduces estimation requirements.

The probability of a value being drawn can be obtained if we know how many multiples of σ separate it from μ

- within σ from μ: $\approx 68\%$
- within 2σ from μ: $\approx 95\%$
- within 3σ from μ: $\approx 99.7\%$



NORMAL DISTRIBUTION

The normal distribution is best suited for data meeting the following minimum requirements:

- strong tendency for the data to take on a central value
- positive, negative deviations from this central value are equally likely
- frequency of the deviations falls off rapidly as we move further away from the central value.

Symmetry of deviations leads to zero **skewness**; low prob. of large deviations from the central value leads to no **kurtosis**.

Its omnipresence in human affairs is linked to the **Central Limit Theorem**.

CENTRAL LIMIT THEOREM

Let $x_1, x_2, \ldots, x_n$ be a **random sample** from any (?) distribution with mean μ and variance σ^2. If the sample observations are **independent** of each other, then the distribution of the average

$$w = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

is **approximately normal** (when $n \to \infty$) with mean and variance

$$\mu_w = \frac{1}{n} E(x_1 + \cdots + x_n) = \mu,\ \ \sigma_w^2 = \frac{1}{n^2} E(x_1 + \cdots + x_n - n\mu)^2 = \frac{1}{n}\sigma^2.$$

The CLT plays an important role in the prevalence of the normal distribution in human affairs.

HOW LARGE IS LARGE?

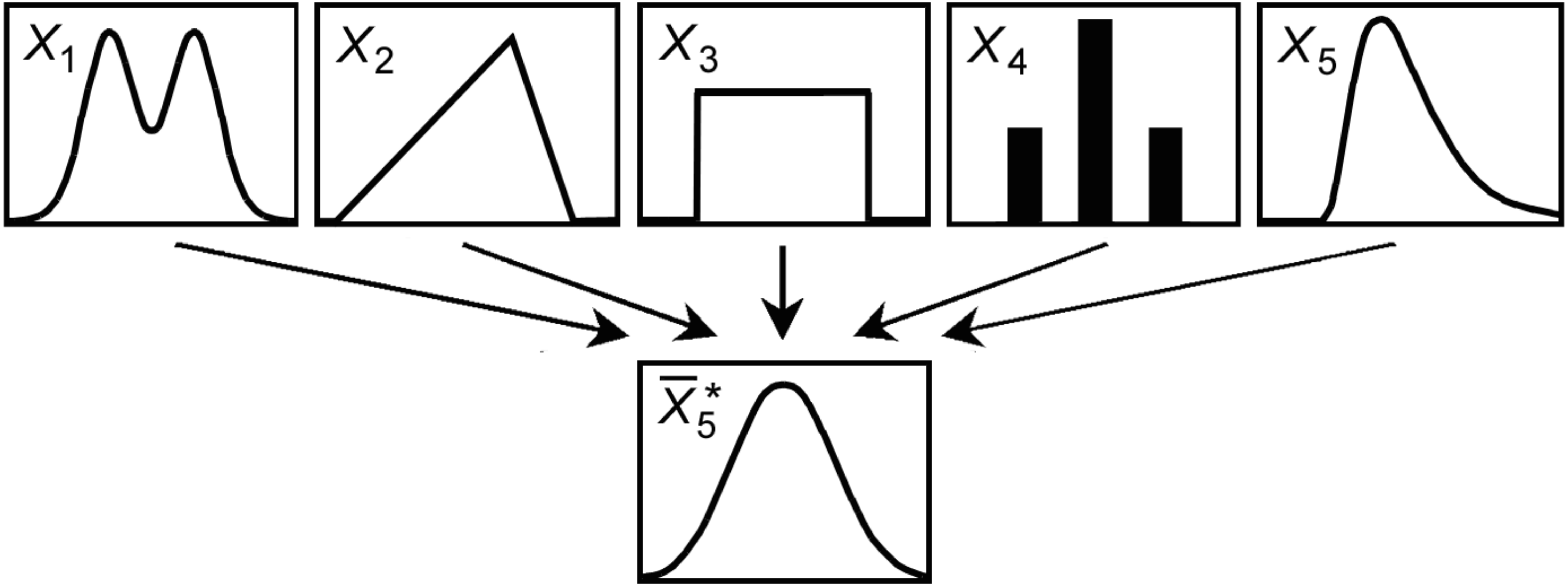
If the underlying population is **normal**, the distribution of the sample mean is also **normal**, no matter the sample size n.

If the underlying population is **approximately symmetric**, the distribution of the sample mean is **approximately normal** for small sample sizes n.

If the sample populations are **skewed** (or **disparate**), the sample size must typically reach 30 before the distribution of the sample mean becomes **approximately normal**.

[https://www.value-at-risk.net/central-limit-theorem]

CENTRAL LIMIT THEOREM IN ACTION



EXERCISES

A large freight elevator can transport a maximum of 9800 lbs. Suppose a load containing 49 boxes must be transported. From experience, the weight of boxes follows a distribution with mean $\mu = 205$ lbs and standard deviation $\sigma = 15$ lbs.

Using R and/or Python, estimate the probability that all 49 boxes can be safely loaded onto the freight elevator and transported.

ESTIMATION

STATISTICAL AND MATHEMATICAL FOUNDATIONS

LEARNING OBJECTIVES

What is estimation, in a statistical sense?

What is estimation used for?

What is bias, in a statistical sense?

ESTIMATION

One of the goals of statistics is to try to **understand a large population** on the basis of the information available in a small sample.

In particular, we are interested in the population **parameters**, which are estimated using suitable sample statistics.

For example, we may use the **sample mean** $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ as an estimate for the true **population mean** μ.

ESTIMATION

The **estimator** is a random variable; the **estimate** is a number.

As an another example, the **sample standard deviation** S is an estimator of the true **population standard deviation** σ and the computed value

$$s = \sqrt{\frac{1}{n-1}\sum_{\mathrm{i}=1}^{n}(x_i - \bar{x})^2}$$

of S is an estimate of σ.

An estimator W of ω is **unbiased** if $\mathrm{E}(W) = \omega$.

BASIC MATHEMATICAL CONCEPTS

Let $X_1, \dots, X_n$ be **random variables**, $b_1, \dots, b_n \in \mathbb{R}$, and E, V, Cov be the **expectation**, **variance**, and **covariance** operators, respectively, i.e.:

- $\mathrm{E}(X_i) = \mu_i$
- $\mathrm{Cov}(X_i, X_j) = \mathrm{E}(X_i X_j) - \mathrm{E}(X_i)\mathrm{E}(X_j)$
- $\mathrm{V}(X_i) = \mathrm{Cov}(X_i, X_i) = \mathrm{E}(X_i^2) - \mathrm{E}^2(X_i) = \mathrm{E}(X_i^2) - \mu_i^2 = \sigma_i^2$ and

$$\mathrm{E}\left(\sum_{i=1}^{n} b_i X_i\right) = \sum_{i=1}^{n} b_i \mathrm{E}(X_i) = \sum_{i=1}^{n} b_i \mu_i$$

$$\mathrm{V}\left(\sum_{i=1}^{n} b_i X_i\right) = \sum_{i=1}^{n} b_i^2 \mathrm{V}(X_i) + \sum_{i \neq j} b_i b_j \,\mathrm{Cov}(X_i, X_j)$$

BASIC MATHEMATICAL CONCEPTS

The **bias** of an estimate is the average of the error in the estimate if the study is repeated many times independently under the same conditions.

The **variability** of an estimate is the extent to which the estimate would vary about its average value in the ideal scenario described above.

The **mean square error** of an estimate is a measure of the error that incorporates both elements:

$$\mathrm{MSE}(\hat{\beta}) = \mathrm{V}(\hat{\beta}) + \mathrm{Bias}^2(\hat{\beta}),$$

where $\hat{\beta}$ is an estimator of β.

BASIC MATHEMATICAL CONCEPTS

If the estimate $\hat{\beta}$ is unbiased, $\mathrm{E}\left(\hat{\beta}-\beta\right)=0$, then an approximate **95% confidence interval** (95% CI) for β is given approximately by

$$\hat{\beta}\pm 2\sqrt{\widehat{\mathrm{V}}\left(\hat{\beta}\right)},$$

where $\widehat{\mathrm{V}}\left(\hat{\beta}\right)$ is a **sampling design-specific** estimate of $\mathrm{V}\left(\hat{\beta}\right)$.

But what is a 95% CI, exactly?

EXERCISE

The total time to manufacture a specific component is known to follow a normal distribution, for which the mean μ and variance σ^2 are not known. In an experiment, 10 components are manufactured; the sample time is given as following:

	1	2	3	4	5	6	7	8	9	10
Time	63.8	60.5	65.3	65.7	61.9	68.2	68.1	64.8	65.8	65.4

What are the best estimates for μ and σ^2? Provide a 95% CI for μ.

BAYES' THEOREM

STATISTICAL AND MATHEMATICAL FOUNDATIONS

LEARNING OBJECTIVES

What is a conditional probability and when is it useful?

What are some mathematical rules that govern probability?

What is Bayes' Theorem and when is it useful?

CONDITIONAL PROBABILITIES

We are often interested in the likelihood of an event occurring **given that another has occurred**.

Examples include:

- the probability that a train arrives on time given that it left on time
- the probability that a PC crashes given the operating system installed
- the probability that a bit is transmitted over a channel is received as a 1 given that the bit transmitted was a 1
- the probability that a website is visited given its number of in-links

Questions of this type are handled using conditional probability.

CONDITIONAL PROBABILITIES

A **conditional probability** is the probability of an event taking place given that another event occurred.

The conditional probability of A given B, $P(A|B)$, is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

The probability that two events A and B both occur is obtained by applying the multiplication rule:

$$P(A \cap B) = P(B)\ P(A|B) = P(A)\ P(B|A)$$

CONDITIONAL PROBABILITIES

Example (a classic): a family has two children (not twins). What is the probability that the youngest child is a girl given that at least one of the children is a girl? Assume that boys and girls are equally likely to be born.

Solution: Let A and B be the events that the youngest child is a girl and that at least one child is a girl, respectively:

$$A = \{GG, BG\}, \qquad B = \{GG, BG, GB\}$$

Then $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{2}{3}$ (not ½, as one might naively assume).

RULES OF PROBABILITY

Let I denote relevant background information; X, Y, Y_k denote propositions, and $-X$ denote the proposition that X is false.

The **plausibility** of X given I is denoted by $P(X|I)$, ranging from 0 (false) to 1 (true).

Sum Rule: $P(X|I) + P(-X|I) = 1$

Product Rule: $P(X, Y|I) = P(X|Y, I)\times P(Y|I)$

Bayes' Theorem: $P(X|Y, I)\times P(Y|I) = P(Y|X, I)\times P(X|I)$

Marginalization Rule: $P(X|I) = \sum P(X, Y_k|I)$, where $\{Y_k\}$ are exhaustive, disjoint

BAYES' THEOREM

The sum rule and the product rules are the **basic rules of probability**.

Bayes' Theorem and the **Marginalization Rule** are simple corollaries of these basic rules.

Bayes' Theorem is sometimes written is a slightly different form

$$P(X|Y,I) = \frac{P(Y|X,I) \times P(X|I)}{P(Y|I)}$$

BAYES' THEOREM

Set-up: assume that an experiment has been conducted to determine the degree of validity of a particular hypothesis, and that experimental data has been collected.

The central data analysis question: given everything that was known *prior* to the experiment, does the collected data support (or invalidate) the hypothesis?

Throughout, let X denote the proposition that the hypothesis in question is true, let Y denote the proposition that the experiment yielded the actual observed data, let I denote (as always) the relevant background information.

BAYES’ THEOREM

Central data analysis question (reprise):

What is the value of $P(\text{hypothesis is true} \mid \text{observed data}, I)$?

Problem: this is nearly always impossible to compute directly.

Solution: using Bayes' Theorem,

$$P(\text{hypothesis} \mid \text{data}, I) = \frac{P(\text{data} \mid \text{hypothesis}, I) \times P(\text{hypothesis} \mid I)}{P(\text{data} \mid I)},$$

it may be that the terms on the right are easier to compute.

BAYES' THEOREM

In the vernacular: the probability

- $P(\text{hypothesis} \mid I)$ of the hypothesis being true prior to the experiment is the **prior**
- $P(\text{hypothesis} \mid \text{data}, I)$ of the hypothesis being true once the experimental data is taken into account is the **posterior**
- $P(\text{data} \mid \text{hypothesis}, I)$ of the experimental data being observed assuming that the hypothesis is true is the **likelihood**
- $P(\text{data} \mid I)$ of the experimental data being observed independently of any hypothesis is the **evidence**

A given hypothesis includes a (potentially implicit) model which can be used to compute or approximate the **likelihood**.

BAYES' THEOREM

Determining the **prior** is a source of considerable controversy

- conservative estimates (uninformative priors) often lead to reasonable results
- in the absence of information, go with maximum entropy prior

The **evidence** is harder to compute on theoretical grounds – evaluating the probability of observing data requires access to some model as part of I. Either

- that model was good, so there's no need for a new hypothesis
- that model was bad, so we dare not trust our computation

BAYES' THEOREM

Thankfully, the evidence is rarely required on problems of parameter estimation (although it is crucial for model selection):

- prior to the experiment, there are numerous competing hypotheses
- the priors and likelihoods will differ, but not the evidence
- the evidence is not needed to differentiate the various hypotheses

Bayes' Theorem is often presented as

$$P(\text{hypothesis} \mid \text{data}, I) \propto P(\text{data} \mid \text{hypothesis}, I) \times P(\text{hypothesis} \mid I)$$

or simply as posterior $\propto$ likelihood$\times$prior, that is to say, **beliefs should be updated in the presence of new information**.

EXERCISE

Suppose that a test for a particular disease has a very high success rate. If a patient

- has the disease, the test accurately reports a ’positive’ with probability 0.99;
- does not have the disease, the test accurately reports a ’negative’ with probability 0.95.

Assume further that only 0.1% of the population has the disease. What is the probability that a patient who tests positive does not in fact have the disease?

MATRIX ALGEBRA

STATISTICAL AND MATHEMATICAL FOUNDATIONS

Neo: What is the Matrix?
Trinity: The answer is out there, Neo. It's looking for you, and it will find you if you want it to.

(*The Matrix*, the Wachowski Sisters)

LEARNING OBJECTIVES

What is the main mathematical object involved in linear algebra?

Why are matrices relevant in data science/data analysis?

What are some matrix operations?

LINEAR ALGEBRA

A **matrix** is an important mathematical tool that allows for easy organization of information, simplifies notation, and facilitates the application of algorithms to data.

Most statistical tools require **rectangular** data:

- each column contains a **variable** (feature, field, attribute)

  - indicator, target, question in a survey, etc.

- each row contains an **observation** (case, unit, item)

  - country, survey respondent, subject in an experiment, etc.

- each cell contains a **value** (measurement) for a particular variable and observation

  - GDP per capita for Canada, answer to a specific question, age, etc.

MATRIX OPERATIONS

A matrix is a rectangular grid of **elements** arranged into **rows** and **columns**.

Matrices are often used in algebra to solve for unknown values in linear equations, and in geometry.

Matrix Addition: matrices can be added together (**element-wise**) as long as their **dimensions** are the same (i.e. both matrices have the same number of rows and columns), like so:

$$\begin{bmatrix} 3 & -2 \\ 4 & 1 \end{bmatrix} + \begin{bmatrix} 4 & 6 \\ 8 & 3 \end{bmatrix} = \begin{bmatrix} 7 & 4 \\ 12 & 4 \end{bmatrix}$$

MATRIX OPERATIONS

Multiplying a Matrix by a Scalar: a matrix of any dimension can be multiplied by a scalar by multiplying each element by the scalar

$$-1\times\begin{bmatrix}2 & 1\\ 3 & -5\\ 4 & 6\end{bmatrix}=\begin{bmatrix}-2 & -1\\ -3 & 5\\ -4 & -6\end{bmatrix}$$

Multiplying Matrices: two matrices A and B can be multiplied if their dimensions are **compatible** (i.e.,$\dim(A)=n\times p$and $\dim(B)=p\times k$). The product $C=AB$ is such that $\dim(C)=n\times k$.

MATRIX OPERATIONS

The element in the *i*th row and *j*th column of the product $C = AB$ is given by

$$c_{i,j} = a_{i,1}b_{1,j} + \cdots + a_{i,p}b_{p,j}$$

For 2×2 matrices, this reduces to

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \times \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} ae + bg & af + bh \\ ce + dg & cf + dh \end{bmatrix}$$

For instance,

$$A = \begin{bmatrix} 4 & 2 & 1 \\ 3 & 0 & 5 \end{bmatrix}, B = \begin{bmatrix} -2 \\ 3 \\ 0 \end{bmatrix} \Longrightarrow AB = \begin{bmatrix} 4\times(-2) + 2\times3 + 1\times0 \\ 3\times(-2) + 0\times3 + 5\times0 \end{bmatrix} = \begin{bmatrix} -2 \\ -6 \end{bmatrix}$$

MATRIX OPERATIONS

Transposing a Matrix: swapping the rows and the columns of a matrix is called **transposing** the matrix – it's denoted with a 'T':

$$\begin{bmatrix} 6 & 0 & -2 \\ 2 & 1 & 3 \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} 6 & 2 \\ 0 & 1 \\ -2 & 3 \end{bmatrix}$$

When applied to a data frame, transposing has the effect of interchanging the role of cases and observations.

For square matrices of size n (i.e. $\dim = n \times n$), there are two special matrices: the **null matrix** 0_n (consisting only of zeroes), and the **identity matrix** I_n (diagonal entries are 1, all others 0).

MATRIX OPERATIONS

For square matrices, two quantities often end up playing a fundamental role: the **trace** and the **determinant**.

The **trace** is the sum of the elements on the main diagonal:

$$\operatorname{tr}\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} = a_{11} + a_{22} + \cdots + a_{nn}$$

MATRIX OPERATIONS

The **determinant** can be computed recursively. Let A be $n \times n$.

1. For $n = 1$, $\det[\, a \,] = a$;

2. For $n = 2$, $\det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = a_{11}a_{22} - a_{12}a_{21}$

3. For a general n, let $D_{i,j}(A)$ be the determinant of the $(n-1)\times(n-1)$ matrix obtained by deleting the i^{th} row and the j^{th} column of A. The **Laplace expansion** of $\det A$ along the first column is

$$(-1)^{1+1}a_{11}\, D_{1,1}(A) + (-1)^{2+1}a_{21}\, D_{2,1}(A) + \cdots + (-1)^{j+1}a_{j1}\, D_{j,1}(A) + \cdots + (-1)^{n+1}a_{n1}\, D_{n,1}(A).$$

MATRIX OPERATIONS

The determinant can be expanded along any row/column without changing its value:

$$\det\begin{bmatrix}1 & 0 & -2\\ 4 & -2 & 6\\ 10 & 8 & 0\end{bmatrix} = 1 \times \det\begin{bmatrix}-2 & 6\\ 8 & 0\end{bmatrix} - 0 \times \det\begin{bmatrix}4 & 6\\ 10 & 0\end{bmatrix} + (-2) \times \det\begin{bmatrix}4 & -2\\ 10 & 8\end{bmatrix} = -152$$

or

$$\det\begin{bmatrix}1 & 0 & -2\\ 4 & -2 & 6\\ 10 & 8 & 0\end{bmatrix} = -0 \times \det\begin{bmatrix}4 & 6\\ 10 & 0\end{bmatrix} + (-2) \times \det\begin{bmatrix}1 & -2\\ 10 & 0\end{bmatrix} - 8 \times \det\begin{bmatrix}1 & -2\\ 4 & 6\end{bmatrix} = -152$$

and

$$\operatorname{tr}\begin{bmatrix}1 & 0 & -2\\ 4 & -2 & 6\\ 10 & 8 & 0\end{bmatrix} = 1 + (-2) + 0$$

MATRIX OPERATIONS

The determinant is linked to the **inverse** of a matrix.

In number arithmetic every number $a \neq 0$ has an inverse b written as a^{-1} or ${}^{1}/_{a}$ such that $ba = ab = 1$. Similarly a square matrix A may have an inverse $B = A^{-1}$ where $AB = BA = I_n$.

Miscellaneous:

- Non-square matrices do not possess inverses.
- Not all square matrices have an inverse (only those with $\det(A) \neq 0$).
- A matrix which has an inverse is said to be **non-singular**.

MATRIX OPERATIONS

If $ad - bc \neq 0$ then the matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ has a (unique) inverse:

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

For $n > 2$, other computation methods exist, such as **Gaussian elimination:** if a sequence of row operations ($yR_j + xR_i \to R_j, R_j \leftrightarrow R_i$) applied to a square matrix A reduce it to the identity matrix I of the same size, then the same sequence of operations applied to I reduces it to A^{-1}.

MATRIX OPERATIONS

If we cannot reduce A to I then A^{-1} does not exist. This will become evident by the appearance of a row of zeros. There is no unique route from A to I and it is experience which selects the optimal route.

It is more efficient to do the two reductions simultaneously;

$$[A|I] = \left[\begin{array}{ccc|ccc} 1 & 3 & 3 & 1 & 0 & 0 \\ 1 & 4 & 3 & 0 & 1 & 0 \\ 2 & 7 & 7 & 0 & 0 & 1 \end{array}\right] \xRightarrow[]{\substack{R_2 - R_1 \to R_2 \\ R_3 - 2R_1 \to R_3}} \left[\begin{array}{ccc|ccc} 1 & 3 & 3 & 1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 1 & 0 \\ 0 & 1 & 1 & -2 & 0 & 1 \end{array}\right]$$

$$\xRightarrow[]{\substack{R_1 - 3R_2 \to R_1 \\ R_3 - R_2 \to R_3}} \left[\begin{array}{ccc|ccc} 1 & 0 & 3 & 4 & -3 & 0 \\ 0 & 1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & -1 & -1 & 1 \end{array}\right] \xRightarrow[]{R_1 - 3R_3 \to R_1} \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 7 & 0 & -3 \\ 0 & 1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & -1 & -1 & 1 \end{array}\right] = [I|A^{-1}]$$

EXERCISES

In R, construct 3×3 square matrices A, B, C and compute the following:

- $A + B, BC, CB, A^{\mathrm{T}}, CA^{\mathrm{T}}$
- $\mathrm{tr}(A), \mathrm{tr}(3A), \mathrm{tr}(C), \mathrm{tr}(-C), \mathrm{tr}(3A - C)$
- $\det(A), \det(A^{\mathrm{T}}), \det(B), \det(C), \det(BC)$
- A^{-1}, B^{-1}, C^{-1}, if the respective determinants are $\neq 0$
- $\det(A^{-1})$, $\det(B^{-1})$, $\det(C^{-1})$, if the respective matrices are invertible

Can you infer rules from these computations?

EIGENVALUES AND EIGENVECTORS

STATISTICAL AND MATHEMATICAL FOUNDATIONS

LEARNING OBJECTIVES

What is an eigenvalue?

What is an eigenvector?

What is a use case for these mathematical concepts?

EIGENVECTORS AND EIGENVALUES

An **eigenvector** of a matrix A is a vector $\boldsymbol{v} \neq \mathbf{0}$ such that, for some scalar λ, $A\boldsymbol{v} = \lambda\boldsymbol{v}$.

The value λ is called an **eigenvalue** of A associated with $\boldsymbol{v}$.

The eigenvalues of an $n\times n$ matrix A satisfy $\det(A - \lambda I_n) = 0$. The left-hand side is a polynomial in λ, and is called the **characteristic polynomial** of A, denoted by $p_A(\lambda)$.

To find the eigenvalues of A, we find the roots of $p_A(\lambda)$.

EXAMPLE

Let $A = \begin{bmatrix} 2 & -4 \\ -1 & -1 \end{bmatrix}$. Then $p_A(\lambda) = \det(A - \lambda I) = (\lambda - 3)(\lambda + 2)$. Thus, $\lambda_1 = 3$ and $\lambda_2 = -2$ are the eigenvalues of A.

To find eigenvectors corresponding to an eigenvalue λ, we solve the system of linear equations given by $(A - \lambda I)\boldsymbol{v} = 0$.

Let's find the eigenvectors corresponding to $\lambda_1 = 3$, by solving

$$(A - 3I)\boldsymbol{v} = \begin{bmatrix} 2-3 & -4 \\ -1 & -1-3 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

EXAMPLE

This yields the following equations:

$$-v_1 - 4v_2 = 0 \ , \qquad -v_1 - 4v_2 = 0$$

If we let $v_2 = t$, then $v_1 = -4t$, and so all eigenvectors corresponding to $\lambda_1 = 3$ are multiples of $\begin{bmatrix} -4 \\ 1 \end{bmatrix}$.

A similar computation shows that all eigenvectors corresponding to $\lambda_2 = -2$ are multiples of $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

EIGEN-DECOMPOSITION

If an $n \times n$ matrix A has n linearly independent eigenvectors, then A may be **decomposed** in the following manner:

$$A = B\Lambda B^{-1},$$

where Λ is a diagonal matrix whose diagonal entries are the eigenvalues of A and the columns of B are the corresponding eigenvectors of A.

EXAMPLE

We have seen that the eigenvalues of $A = \begin{bmatrix} 2 & -4 \\ -1 & -1 \end{bmatrix}$ are $\lambda_1 = 3$ and $\lambda_2 = -2$, and that the corresponding eigenvectors are $\begin{bmatrix} -4 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

Thus, $\Lambda = \begin{bmatrix} 3 & 0 \\ 0 & -2 \end{bmatrix}$, $B = \begin{bmatrix} -4 & 1 \\ 1 & 1 \end{bmatrix}$, and

$$A = \begin{bmatrix} -4 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & -2 \end{bmatrix} \begin{bmatrix} -4 & 1 \\ 1 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} -4 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & -2 \end{bmatrix} \frac{1}{-4\times 1 - 1\times 1} \begin{bmatrix} 1 & -1 \\ -1 & -4 \end{bmatrix}$$

$$= \begin{bmatrix} -4 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & -2 \end{bmatrix} \begin{bmatrix} -1/5 & 1/5 \\ 1/5 & 4/5 \end{bmatrix}$$

EXERCISES

Compute the eigen-decomposition of the matrices A, B, C you constructed in the previous module.

REGRESSION

STATISTICAL AND MATHEMATICAL FOUNDATIONS

LEARNING OBJECTIVES

What is regression modelling?

What are some types of regression modeling?

When is regression modeling useful?

REGRESSION MODELING

The most common data modeling methods are regressions, both **linear** and **logistic**

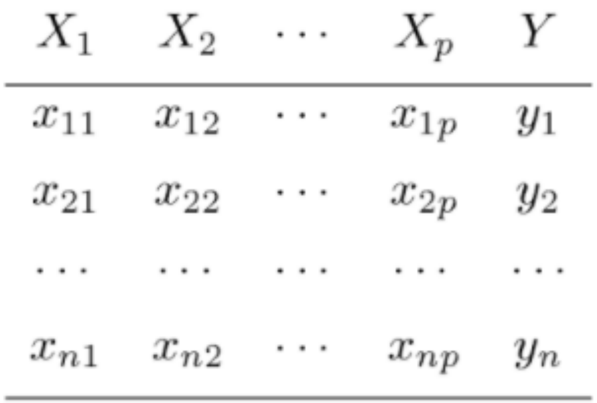
- ~90% of real data applications end up using a simple regression as their final model, typically after very careful data preparation, encoding, and creation of variables.

There are several reasons for their frequent use:

- generally straightforward to understand and to train
- mean square error (MSE) objective function has a closed-form linear solution
- system of equations can usually be solved through matrix inversion or linear manipulation

REGRESSION MODELING

The data structure of a general modeling task is represented by →



X_1	X_2	$\cdots$	X_p	Y
x_{11}	x_{12}	$\cdots$	x_{1p}	y_1
x_{21}	x_{22}	$\cdots$	x_{2p}	y_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_{n1}	x_{n2}	$\cdots$	x_{np}	y_n

We consider p independent variables X_i to try to predict the dependent variable Y.

In order to simplify the discussion in the following, we introduce the matrix notation $\boldsymbol{X}[n\times p], \boldsymbol{Y}[n\times 1], \boldsymbol{\beta}[p\times 1]$, where n is the # of observations and p is the # of independent variables.

LINEAR REGRESSION

The basic assumption of linear regression is that the dependent variable y can be **approximated** by a linear combination of the independent variables as follows:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\beta \in \mathbb{R}^p$is to be determined based on the training set, and for which

$$\mathrm{E}(\boldsymbol{\varepsilon}|\boldsymbol{X}) = 0, \qquad \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T|\boldsymbol{X}) = \sigma^2 I.$$

Typically, the errors are also assumed to be normally distributed, that is :

$$\boldsymbol{\varepsilon}|\boldsymbol{X} \sim N(0, \sigma^2 I).$$

LINEAR REGRESSION

If $\hat{\beta}_i$ is the estimate of the true coefficient β_i, the **linear regression** model associated with the data is

$$\widehat{\boldsymbol{Y}}(\boldsymbol{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

In matrix form, the regression problem requires a solution $\widehat{\boldsymbol{\beta}}$ to the **normal equation** $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{Y}$.

When the symmetric positive definite matrix $\boldsymbol{X}^T\boldsymbol{X}$ is invertible, the fitted coefficient is simply $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{Y})$. Note that $\boldsymbol{X}^T\boldsymbol{X}$ is a $p \times p$ matrix, which makes the inversion "easier" to compute, relatively speaking, when n is large.

GENERALIZED LINEAR REGRESSION

Generalized linear models (GLMs) extend linear statistical models by accommodating response variables with **non-normal** conditional distributions.

Except for the **error structure**, a GLM is essentially the same as for a linear model:

$$Y_i \sim \text{ some distribution with mean } \mu_i, \text{ where } g(\mu_i) = x_i^T\beta$$

A GLM therefore consists of three parts:

- a **systematic** component $x_i^T\beta$
- a **random** component – specified distribution for Y_i
- a **link** function g

GENERALIZED LINEAR REGRESSION

We could specify **any** distribution for the outcome variable Y...

- but the mathematics of GLM work nicely only for the **exponential family** of distributions (most common statistical distributions fall into this family: such as the normal, binomial, Poisson, gamma, and others).

Linear regression is an example of GLM:

- systematic component: $x_i^T \beta$
- random component: $Y_i \sim N(\mu_i , \sigma^2)$
- link: $g(\mu) = \mu$, the identity link

EXAMPLE

In the early stages of an epidemic, the rate at which new cases occur increases exponentially through time.

If μ_i is the expected number of new cases on day t_i, a model taking the form

$$\mu_i = \gamma \exp(\delta t_i)$$

might be appropriate. If we take the log of both sides, we get

$$\underset{\text{link}}{\log(\mu_i)} = \log(\gamma) + \ \delta t_i = \beta_0 + \beta_1 t_i = \underset{\text{systematic component}}{(1, t_i)^T(\beta_0, \beta_1)}.$$

Furthermore, since the we measure the number of new cases (a count), the **Poisson** distribution could be a reasonable choice.

random component

ADVANTAGES OF GLM

No need to transform Y to have a normal distribution

Choice of link is **separate** from the choice of random component

- more modeling flexibility

If link produces **additive effects**, no need for constant variance

Models are fitted via ML estimation

- optimal properties of the estimators

Inference tools and **model checks** apply to other GLMs

- Wald ratio test, likelihood ratio test, deviance, residuals, confidence intervals, etc.

See `PROC GENMOD` in SAS, or `glm()` in R

EXERCISE

An auto part is manufactured by a company once a month in lots that vary in size as demand fluctuates. The data below represent observations on lot size (y), and number of employee-hours of labor (x) for 10 recent production runs.

Fit a simple regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where $\mathrm{E}(\varepsilon_i) = 0$, $\mathrm{E}\left(\varepsilon_i \varepsilon_j\right) = 0$ for $i \neq j$, and $\mathrm{V}(\varepsilon_i)= \sigma^2$ if the observations are:

$$\boldsymbol{Y} = [73,50,128,170,87,108,135,69,148,132]^{\mathrm{T}},$$

$$\boldsymbol{x} = [30,20,60,80,40,50,60,30,70,60]^{\mathrm{T}}.$$

OPTIMIZATION

STATISTICAL AND MATHEMATICAL FOUNDATIONS

LEARNING OBJECTIVES

What is optimization?

When is optimization useful?

What is a cost function?

Why are minima and maxima relevant to optimization?

What are techniques that can be used to carry out optimization?

OPTIMIZATION

Suppose we have a **cost** (objective) function $f\colon \mathbb{R}^n \to \mathbb{R}$ to **optimize** (the maximum likelihood function of linear regression, for instance).

Seeking a maximum for f is equivalent to seeking a minimum for $-f$.

The aim is to find parameter values $\boldsymbol{x}$ that minimize this function:

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}} f(\boldsymbol{x})$$

The cost function could be subjected to a number of constraints

$$c_i(\boldsymbol{x}) = 0, i = 1, \ldots, m;\ c_j(\boldsymbol{x}) \geq 0, j = 1, \ldots, k;\ \boldsymbol{x} \in \Omega \subseteq \mathbb{R}^n.$$

OPTIMIZATION

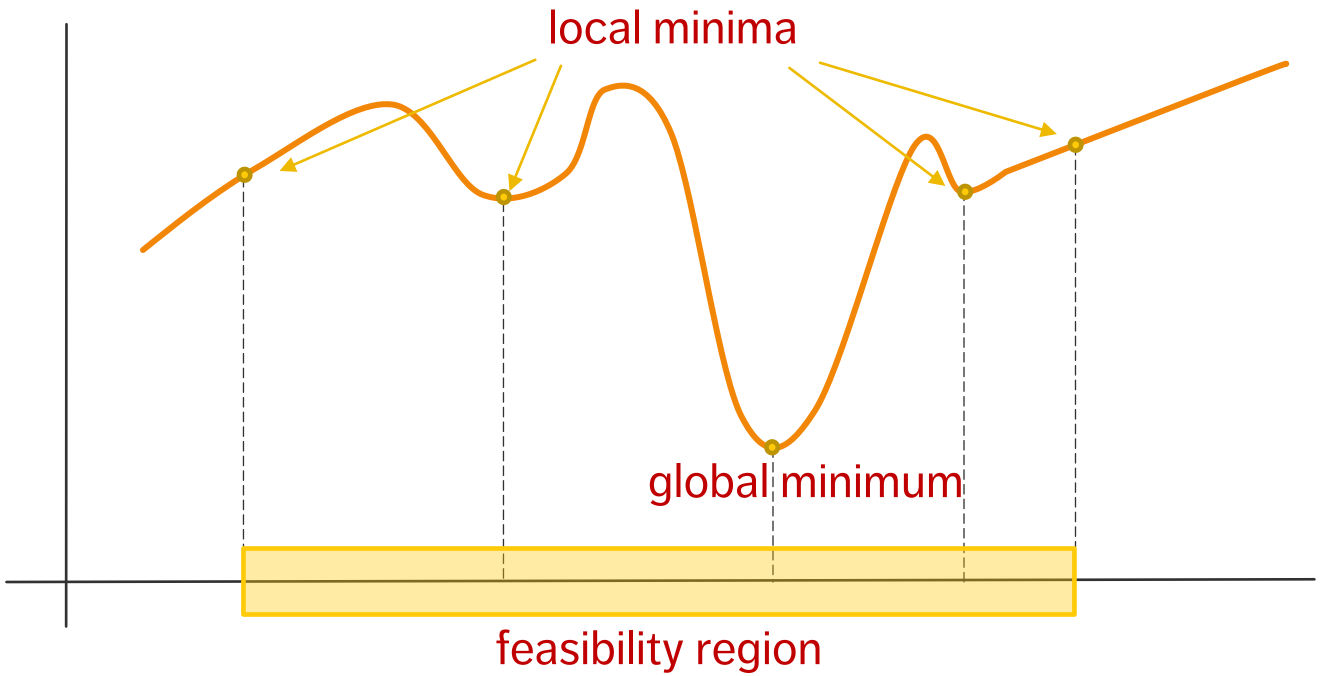
The optimization problem can be viewed as a **decision problem** that involves finding the “best” vector $\mathbf{x}$ over all possible vectors in $\Omega \subseteq \mathbb{R}^n$.

This vector is called the **minimizer** of f over Ω. There may be multiple minimizers, or none.

If $\Omega = \mathbb{R}^n$, then we refer to the problem as an **unconstrained** optimization problem.

In general, this is not a trivial problem (consult the literature).

TYPE OF MINIMA



In many instances, optimization is a **numerical** endeavour. Which of the minima is found depends on the algorithm‘s **starting point**.

GOLDEN SECTION METHOD

The **golden section search** is a technique for finding the extremum (minimum or maximum) of a strictly unimodal function by successively narrowing the range of values inside which the extremum is known to exist.

The technique derives its name from the fact that the algorithm maintains the function values for triples of points whose distances form a **golden ratio**.

GOLDEN SECTION METHOD

Let $[a, b]$ be the interval of the current bracket (i.e. the optimizer resides in $[a, b]$), and assume $f(a), f(b)$ have already have been computed. Denote $\varphi = (1+\sqrt{5})/2$.

1. Let $c = b - \frac{(b-a)}{\varphi}$, $d = a + \frac{(b-a)}{\varphi}$;
2. If $f(c), f(d)$ are not available, compute them;
3. If $f(c) < f(d)$ (to find a min – to find a max, reverse the order) then move the data: $\big(b, f(b)\big) \leftarrow \big(d, f(d)\big)$ and $(d, f(d)) \leftarrow (c, f(c))$ and update $c = b - (a - b)/\varphi$ and $f(c)$;
4. Otherwise, move the data $\big(a, f(a)\big) \leftarrow \big(c, f(c)\big)$ and $(c, f(c)) \leftarrow (d, f(d))$ and update $d = a + (b - a)/\varphi$ and $f(d)$;
5. The interval $[c, d]$ brackets the optimizer. Continue until tolerance is reached.

NEWTON'S METHOD

In calculus, we learn that a function $f: \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$ which is sufficiently well behaved reaches its max/min either at a **critical point** (i.e. where $\nabla f = \mathbf{0}$) or on the **domain boundary** $\partial\Omega$.

Thus, to identify candidate optimizers, we must be able to solve general systems of the form $g(\boldsymbol{x}) = \mathbf{0}$.

Newton's Method is a powerful method for finding roots of functions.

NEWTON'S METHOD

For $n = 1$, the algorithm is shown below (it is quite similar in the general case).

Let $x = c$ be an (unknown) zero of a differentiable function f in an open interval containing c.

1. make an initial approximation x_1 "close" to c
2. determine a new approximation using the formula $x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$.
3. If $|x_2 - x_1|$ is less than the desired accuracy (which needs to be specified), x_2 serves as the final approximation. Otherwise, return to step 2. and calculate a new approximation.

EXERCISES

Use the golden section method and Newton's method to find a root of

$$f(x) = e^{-x}\sin(x) \text{ and } g(x) = x\ln(x).$$

REFERENCES

STATISTICAL AND MATHEMATICAL FOUNDATIONS

REFERENCES

Wu, J., Coggeshall, S. [2012], *Foundations of Predictive Analytics,* CRC Press.

Bruce, P., Bruce, A. [2017], *Practical Statistics for Data Scientists, 50 Essential Concepts*, O'Reilly.

Jaynes, E.T. [2003], *Probability Theory: the Logic of Science*, Cambridge

https://ece.uwaterloo.ca/~dwharder/NumericalAnalysis/11Optimization/newton/

https://www.math.ucdavis.edu/~thomases/W11_16C1_lec_3_11_11.pdf

https://web.as.uky.edu/statistics/users/pbreheny/760/S13/notes/1-24.pdf

https://socialsciences.mcmaster.ca/jfox/Courses/SPIDA/GLMs-notes.pdf

https://onlinecourses.science.psu.edu/stat504/node/216

http://stattrek.com/regression/regression-example.aspx?Tutorial=AP

http://www.stat.ucla.edu/~nchristo/introeconometrics/introecon_matrix_simple_regr.pdf

REFERENCES

http://www.personal.soton.ac.uk/jav/soton/HELM/workbooks/workbook_30/30_3_lu_decomposition.pdf

https://www.math.hmc.edu/calculus/tutorials/eigenstuff/

https://people.duke.edu/~ccc14/sta-663/LinearAlgebraMatrixDecompWithSolutions.html

https://www.georgebrown.ca/uploadedFiles/TLC/_documents/Basic%20Matrix%20Operations.pdf

https://bgsu.instructure.com/courses/901773/pages/p5-learning-using-bayes-rule?module_item_id=6367315

http://www4.stat.ncsu.edu/~bmasmith/ST371S11/Conditional-Probability-and-Independence.pdf

https://www2.isye.gatech.edu/~brani/isyebayes/bank/handout1.pdf

Joint probability distribution on Wikipedia

Multivariate normal probability distribution on Wikipedia