TMA4267 Linear Statistical Models V2017 [L1]

Introduction to the course
Part 1: Multivariate random variables, and the multivariate normal distribution

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: January 10, 2017

TMA4267 Linear Statistical Models

- Statistics, linear statistical models and movie recommender systems.
- Learning outcome.
- TMA4267 core and parts.
- Background knowledge in probability and statistical inference.
- TMA4267 course information.
- Voting and questionnaire.
- Part 1: Multivariate RVs and the multivariate normal distribution.

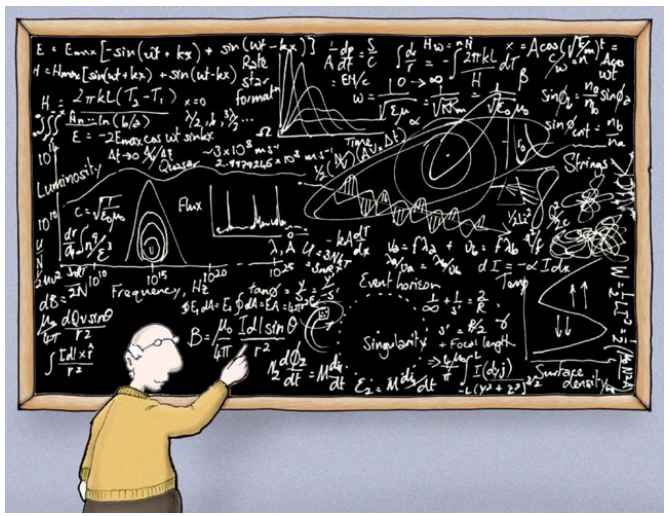
What is Statistics?

- The true foundation of theology is to ascertain the character of God.
- It is by the aid of *Statistics* that law in the social sphere can be ascertained and codified,
- and, certain aspects of the character of God hereby revealed.
- The study of statistics is thus a *religious service.*

Florence Nightingale (1820-1910). Quotation from "Games, Gods and Gambling: A History of Probability and Statistical Ideas" by F. N. David.

What is Statistics?

- The goal of Statistics is to expand our knowledge based on collection and analysis of empirical data.
- Two branches:
  - Probability: the mathematical study of the probability of random events.
  - Statistical Inference: models and methods for collecting, describing, analysing and interpreting numerical data.



Drawing taken from http://www.nearingzero.net - now at http://www.lab-initio.com/

Word cloud: Probability

Word cloud: Statistical Inference

Linear Statistical Models

Simple linear regression (height of child explained by mid-parent height):

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Multiple linear regression (also include other explanatory variables):

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

The multiple linear regression model is our linear statistical model!
So, why is this course not called "Regression"?
We include theory that focus on mathematical understanding: multivariate random variables, the multivariate normal distribution, projections, idempotent matrices, hypothesis tests, design of experiments, ...

Recommender systems

- Recommender systems are a subclass of information filtering system that seek to predict the 'rating' or 'preference' that a user would give to an item.
- Recommender systems have become extremely common in recent years, and are applied in a variety of applications. The most popular ones are probably movies, music, news, books, research articles, search queries, social tags, and products in general. However, there are also recommender systems for experts, collaborators, jokes, restaurants, financial services,life insurance, persons (online dating), and Twitter followers.

Source: Wikipedia: Recommender systems

The Netflix Price: 2006

Text from `http:\www.netflixprice.com`.

- To help customers find those movies, we've developed our world-class movie recommendation system: Cinematch.
- Its job is to predict whether someone will enjoy a movie based on how much they liked or disliked other movies. We use those predictions to make personal movie recommendations based on each customer's unique tastes. And while Cinematch is doing pretty well, it can always be made better.
- We provide you with a lot of anonymous rating data, and a prediction accuracy bar that is 10% better than what Cinematch can do on the same training data set. (Accuracy is a measurement of how closely predicted ratings of movies match subsequent actual ratings.)
- Remark: At this point in time DVDs were sent to customers by mail - this was before the age of online streaming.

The prize was awarded the team *BellKor's Pragmatic Chaos* in 2009.

Cinematch

The Cinematch recommender system: *use statistical linear models with a lot of data conditioning.* I have not found any other information on the algorithm online.

"Simple" linear suggestion:
(predicted score on movie for person)=
(some overall score for this movie)+
(some overall score used by this person)+
(similarity of this movie with other movie this person has seen)*
(how much this person liked that movie)+
the same for all the movies this person has rated+
error term.

The Netflix Price: Training data

- The training data set consists of more than 100 million ratings from over 480 thousand randomly-chosen, anonymous customers on nearly 18 thousand movie titles.
- The ratings are on a scale from 1 to 5 (integral) stars. The date of each rating and the title and year of release for each movie are provided.
- No other customer or movie information is provided. No other data were employed to compute Cinematch's accuracy values used in this Contest.

Text from `http:\www.netflixprice.com`.

The Netflix Price: Test data

- A qualifying test set is provided containing over 2.8 million customer/movie id pairs with rating dates but with the ratings withheld.
- Eligible algorithms must provide predictions for all the withheld ratings for each customer/movie id pair in the qualifying set.
- The qualifying set is divided into two disjoint subsets containing randomly selected pairs from the qualifying set. The assignment of pairs to these subsets is not disclosed.
  - The Site will score each subset by computing the square root of the averaged squared difference between each prediction and the actual rating (the root mean squared error or "RMSE") in the subset, rounded to the nearest .0001.
  - The RMSE for the first "quiz" subset will be reported publicly on the Site,
  - the RMSE for the second "test" subset will not be reported publicly but will be employed to qualify a submission as described below.

Text from `http:\www.netflixprice.com`.

The winning algorithm: lessons to learn

Bell, Koren and Volinsky (2010): *All Together Now: A Perspective on the Netflix Price*, Chance, 23, p. 24-29.

- Most entries into the competition looked at the problem as a set of algorithms – focus on prediction rather than on understanding what *drives* the preditions.
- Complex models are prone to over fitting or matching small details rather than the big picture, especially where data are scarce (importance of cross-validation).
- The final model is an *ensemble model* combining many different prediction models (at least more than 100), including nearest neighbour methods, latent factor models, neural networks, weighting determined by ridge regression.
- The winning model was never implemented by Netflix, partly due to implementation issues - but also due to the increase of available data after "sending DVDs by mail" was replaced by online streaming.

Read more: Link to talk with interesting points raised.

TMA4267 Linear statistical methods
Learning outcome, Knowledge

- The student has strong theoretical knowledge about the most popular statistical models and methods that are used in science and technology, with emphasis on regression-type statistical models.
- The statistical properties of the multivariate normal distribution are well known to the student, and the student is familiar with the role of the multivariate normal distribution within linear statistical models.

TMA4267 Linear statistical methods
Learning outcome, Skills

- The student knows how to design an experiment and
- how to collect informative data of high quality to study a phenomenon of interest.
- Subsequently, the student is able to choose a suitable statistical model,
- apply sound statistical methods, and
- perform the analyses using statistical software.
- The student knows how to present the results from the statistical analyses, and how to draw conclusions about the phenomenon under study.

TMA4267: Parts

- Part 1: Multivariate RVs and the multivariate normal distribution [week 2-5].
  - Data consists of simultaneous measurements on many variables: we work with *random vectors and random matrices*.
  - There is a strong connection between the *multivariate normal distribution* and the classical linear model.
- Part 2: The classical linear model [week 6-9]
  - We want to understand the relationship between many variables: with focus on linear relationships through the *classical linear model* (multiple linear regression).
- Part 3: Hypothesis tests and analysis of variance [week 9-11]
  - Is there and association between a response and an explanatory variable? Does a response vary between treatment groups?
- Part 4: Design of Experiments [week 12-13+project]
  - If we want to collect data, we need to do know how to *design an experiment*.

Do you know this?

Recommended background: TMA4240/TMA4245 Statistics.

- Probability: (continuous) random variables (RV), probability distribution function (pdf), cumulative distribution function (cdf), mean E, variance Var, covariance Cov, correlation Corr, transformation formula, momentgenerating function (MFG), normal, chi-square and t-distributions.
- Inference: population and sample philosophy, parameter estimation, confidence interval, hypothesis test, p-value, simple linear regression.
- Linear methods: vector and matrix algebra (trace, determinant, eigenvalues/vectors), real vector spaces, orthogonality, spectral decomposition.

TMA4267 Linear Statistical Models
Course information

`https://innsida.ntnu.no/bb`

- Course information.
- Course material.
- Lectures (and handouts).
- Statistical software.
- Exercises (6 recommended and 4 compulsory).
- Exam (80% of portfolio assessment).

Is this the correct course for you?

Are you afraid that this course have a too strong focus on theory and to little on the practical aspects of statistics? You may also look at at the following similar courses (that is, a second course in statistics, with focus on inference)

- ST2304 Statistical modelling for biology/biotechnology: `https://wiki.math.ntnu.no/st2304/`
- TMA4255 Applied statistics, for all siv.ing. studiprograms (except IndMat): `https://wiki.math.ntnu.no/tma4255/`
- KLMED Medical statistics II: `https://www.ntnu.no/studier/emner/KLMED8005`

Electronic voting

– more than an anonymous show of hands?

For student: check that topics are understood, compare to class, focus on the question asked, while preserving anonymity.

For lecturer: collect data to design sessions that are more contingent.

Software: clicker. math.ntnu.no (single questions), Kahoot! (end-of-lecture sum-up), quiz in Blackboard.

Future studies?

What is your current plan of topic for future studies?

- A: Statistics
- B: Mathematics
- C: Numerics
- D: Other
- E: Don't know

Electronic voting

Use your smart phone, or other devise with internet access and go to **http://clicker.math.ntnu.no/**, and then select TMA4267 as classroom.

Answers

- A: Statistics
- B: Mathematics
- C: Numerics
- D: Other
- E: Don't know

Start voting now!

Part 1: Multivariate random vectors and the multivariate normal distribution

- Härdle and Simar (2015): *Applied Multivariate Statistical Analysis*. Springer.
  - Chapter 2 (p. 53-76): A Short Excursion into Matrix Algebra (partly lectured, manly assumed known).
  - Chapter 3.3 (p. 89-93): Summary statistics.
  - Chapter 4.1-4.5 (p. 117-149): Multivariate Distributions.
  - Chapter 5.1 (p. 183-190): Elementary Properties of the Multinormal.
- Fahrmair, Kneib, Lang and Marx (2013): *Regression*. Springer.
  - Appendix B: Def B.11 (chis q), B.13 (t) , B14 (F), Theorem B.2 and B3.3 (distribution of quadratic forms).

A merged pdf named TMA4267Part1.pdf is available from Bb. Both eBooks and can be downloaded without charge for NTNU students.

The Cork deposit data

- Classical data set from Rao (1948).
- Weigth of bark deposits of $n = 28$ cork trees in $p = 4$ directions (N, E, S, W).

Tree	N	E	S	W
1	72	66	76	77
2	60	53	66	63
3	56	57	64	58
⋮	⋮	⋮	⋮	⋮
28	48	54	57	43

How may we define a random vector in connection to the cork deposit data set?

Hands-on

Let $\boldsymbol{X}_{(2\times 1)}$ have joint pdf (see the 3D-printed figure)

$$f(x_1, x_2) = \frac{1}{2\pi} e^{-\frac{1}{2}(x_1^2 + x_2^2)} \text{ for } -\infty < x_1, x_2 < \infty$$

Find:

1. the marginal distributions $f_1(x_1)$ and $f_2(x_2)$,
2. the conditional distributions $f(x_1 \mid x_2)$ and $f(x_2 \mid x_1)$.
3. What about $F(x_1, x_2)$?

Hint (why?):

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = 1$$

Copula [H4.1, p120]

A two-dimensional copula is a function $C : [0,1]^2 \to [0,1]$ with the following properties:

- For every $u \in [0,1]$: $C(0,u) = C(u,0) = 0$.
- For every $u \in [0,1]$: $C(u,1) = u$ and $C(1,u) = u$.
- For every $(u_1,u_2),(v_1,v_2) \in [0,1] \times [0,1]$ with $u_1 \leq v_1$ and $u_2 \leq v_2$:

$$C(v_1,v_2) - C(v_1,u_2) - C(u_1,v_2) + C(u_1,u_2) \geq 0$$

(The last property is called "2-increasing".)

Remark: this is not part of the core of the course (not suitable as an exam question), but it is a nice concept and you should have heard about it.

Sklar's Theorem [H4.1, p121-122]

Let F be a joint (cumulative) distribution function with marginal distribution functions F_1 and F_2. Then a copula C exists with

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2))$$

for every $x_1, x_2 \in \mathbb{R}$. If F_1 and F_2 are continuous, then C is unique. On the other hand, if C is a copula and F_1 and F_2 are (cumulative) distribution functions, then the function F defined above, is a joint distribution function with marginals F_1 and F_2.

Remark: this is not part of the core of the course (not suitable as an exam question), but it is a nice concept and you should have heard about it.

Bivariate Copulas

Farlie-Gumbel-Morgenstern family

$$C(u,v) = uv + \theta uv(1-u)(1-v),\ \theta \in [-1,1]$$

The only copulas that are polynomial quadratic in u and v, symmetric.

Normal (Gaussian) copulas [H.p141]

$$f(x_1,x_2) = \frac{1}{2\pi}\frac{1}{\sigma_1\sigma_2\sqrt{1-\rho^2}}e^{-\frac{1}{2}Q(x_1,x_2)}$$

$$Q(x_1,x_2) = \frac{1}{1-\rho^2}\left[(\frac{x_1-\mu_1}{\sigma_1})^2 + (\frac{x_2-\mu_2}{\sigma_2})^2 - 2\rho(\frac{x_1-\mu_1}{\sigma_1})(\frac{x_2-\mu_2}{\sigma_2})\right]$$

$$C(u,v) = \int_{-\infty}^{\Phi_1^{-1}(u)}\int_{-\infty}^{\Phi_2^{-1}(v)} f(x_1,x_2)dx_1dx_2$$

Read more? Properties and applications of copulas: A brief survey, Roger B. Nelsen (And same remark as before.)

More on copulas?

We will later in Part 1 (recommended exercise 2) look at data and contour plots from Gaussian copulas, and other copulas poplar in finance - using R and the `copula` library in R.

What have we worked with today?

- A random vector is ... a vector of random variables.
- Joint distribution function.
- From joint distribution function to marginal and conditional distributions.
- Cumulative distribution.
- Independence.
- From marginal cumulative distribution functions to joint using copula.

Next lecture: Mean vector and covariance matrix. You may want to look into how to define a positive definite matrix, how to define eigenvalues/vectors and results for symmetric matrices.