Module 1: INTRODUCTION

TMA4315 Generalized linear models H2023

Bob O'Hara, Department of Mathematical Sciences, NTNU

22.08 [PL] and 23.08 [IL]

Introduction: Aim of this module

- Introduction (to the introduction...)
- What we will teach
- How we will teach
- Where we are going with GLMs
- short presentation of all course modules
- learning outcomes
- student learning styles
- interactive lectures: what, why and how?
- practical details of the course (Blackboard)

Introduction: Aim of this module

- core concept: the exponential family of distributions
- learn about - and use - `R`, `Rstudio`, `R Markdown`, and get familiar with related topics
- get up to speed on R (and writing reports in R markdown) to be able to do the 3*10-points compulsory exercises by doing recommended exercises

Where we are going with GLMs: Expanding the linear regression framework

We will stay with regression (for the whole course) - but make expansions in several directions.

What will not change:

- our target is *a random response* Y_i (from some statistical distribution): continuous, binary, nominal or ordinal, we have
- *fixed covariates (or explanatory variables)* X_i (in a design matrix): quantitative or qualitative, and
- *unknown regression parameters* β.

Expanding the linear regression framework

We will consider relationships between the *conditional mean of* Y_i, $\mathsf{E}(Y_i \mid \mathbf{x}_i) = \mu_i$, and linear combinations of the covariates in *a linear predictor*

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} = \mathbf{x}_i^T \ .$$

We connect η_i and μ through a *link function*:

$$\mu_i = g^{-1}(\eta_i)$$

Expanding the linear regression framework even more

For most of the course we will assume observation pairs $(Y_i, \mathbf{x}_i)$ are independent $i = 1, \ldots, n$, but we will also consider clustered pairs (in Module 7+8: Linear mixed effects models LMM and Generalized linear mixed effects models GLMM).

$$\eta_i = \mathbf{x}_i^T \ + \mathbf{Zu}$$

(notation to be explained: the take-home message is that this is still linear)

Modules

The course is split up into 9 modules, which take 1-2 weeks of class.

Each module has a different theme

The modules - in short

The modules of this course are:

1. Introduction (the module page you are reading now) [week 34]
2. Multiple linear regression (emphasis on likelihood) [week 35-36]
3. Binary regression (binary individual and grouped response) [week 37-38]
4. Poisson and gamma regression (count, non-normal continuous) [week 39-40]
5. GLM in general and quasi likelihood (exponential family, link function) [week 41]
6. Categorical regression and contingency tables [week 43]
7. Linear mixed models (clustered data, repeated measurements) [week 44-45]
8. Generalized mixed effects models [week 46]
9. Discussion and conclusion [week 47]

Common - for all modules

1. Model specification: an equation linking the response and the explanatory variables, and a probability distribution for the response.
2. Estimation of the parameters in the model
3. Checking the adequacy of the model, how well it fits the data.
4. Inference: confidence intervals, hypothesis tests, interpretation of results, prediction of future responses.

Both theoretic derivations and practical analysis in R will be emphasized.

Module 2: Multiple linear regression

PLAN: You recapitulate what you have learned in TMA4267 Linear statistical models, and in the plenary lectures we focus on a three-step model, likelihood theory and formal inference connected to the likelihood. Instead of sums-of-squares of error (MSE, RSS) we will use deviance.

In Compulsory exercise 1 you make your own `mylm` function to perform MLR.

Textbook: Chapter 3 (from TMA4267) and parts of Appendix B4.

Module 3: Binary regression

How can be model a response that is not a continuous variable? Here we look at present/absent, true/false, healthy/diseased.

PLAN: In this module we will study the binary regression, work on parameter estimation and interpretation of parameter estimates using odds, work with both individual and grouped data, test linear hypotheses, look at criteria for model fit and model choice, and discuss overdispersion.

Textbook: 2.3 and 5.1.

Module 4: Poisson and gamma regression

Count data - the number of times and event occurs - is common.
Continuous positive data - like life times, costs and claim sized
Plan We will look at effect a one of more covariates that may work multiplicative on the response and see how we may fit that assuming a Poisson distribution (counts) or gamma regression (continuous) on the log scale of the response.
Textbook: 5.2 and 5.3

Module 5: GLM in general (and quasi likelihood — if time)

We will see that normal, binary, Poisson and gamma regression all have the same underlying features: this leads to a unified framework, and maximum likelihood estimation can be written on a generalized form for all GLMs

Plan Develop the maths for GLMs, including their statistical inference and asymptotic properties of estimators on a common form. Finally, we may expand this to quasi-likelihood models by just specifying mean and variance (not distribution).

This part is rather mathematical - but is built on the findings of modules 1-4.

Textbook: 5.4 and 5.5

Module 6: Categorical regression and contingency tables

Here our response variable has more than two categories, and these categories can either be unordered or ordered.

Plan We will use the multinomial distribution as the distribution for the response, and work mainly with grouped data - that often can be presented in a contingency table. For ordered categories (like the defoliation of trees) we will use a cumulative model, also called a proportional odds model.

Textbook: Chapter 6, and possibly extra material on the Fisher and Chi-square test (if time permits).

Compulsory exercise 2 will cover modules 3-6.

Module 7: Linear mixed effects models

We sometimes have categorical factors with lots of levels, e.g. repeated measures on several subjects

For example, someone might have observed that subjects' reaction times on different days as they reduce the amount of sleep they get. Each subject might have different baseline reactions, nad might also respond differently to sleep.

In linear mixed effects models we assume that the intercepts and slopes are drawn from normal distributions and estimate the variance in these distribution. The model makes observations correlated within subjects.

Plan We will look at different models for clustered and repeated measurement (e.g. over time) using regression with fixed and random effects.

Textbook: 2.4, 7.1, 7.3

Module 8: Generalized linear mixed effects models

We generalize our model in Module 7 - on normal responses - to binary (and possibly Poisson) responses.

Textbook: 7.2, 7.5, 7.7

Compulsory exercise 3 will cover modules 7-8.

Module 9: Discussion and Conclusions

Summarise where we are

Textbook

Textbook: Fahrmeir, Kneib, Lang, Marx (2013): "Regression. Models, Methods and Applications" https://link.springer.com/book/10.1007%2F978-3-642-34333-9 (free ebook for NTNU students). Tentative reading list: main parts of Chapters 2, 3 (repetition), 5, 6, 7, Appendix B.4.

Learning outcomes

Knowledge.

The student can assess whether a generalised linear model can be used in a given situation and can further carry out and evaluate such a statistical analysis. The student has substantial knowledge of generalised linear models and associated inference and evaluation methods. This includes regression models for Gaussian distributed data, logistic regression for binary and multinomial data, Poisson regression and log-linear models for contingency tables.

The student has theoretical knowledge about linear mixed models and generalized linear mixed effects models, and associated inference and evaluation of the models. Main emphasis is on Gaussian and binomial data.

Skills.

The student can assess whether a generalised linear model or a generalized linear mixed model can be used in a given situation, and can further carry out and evaluate such a statistical analysis.

How we will teach

Two principles used in developnig this course

- learning styles
- active learning

Learning styles

Back in 1988 Felder and Silverman devised a taxonomy for learning styles - where four different axis are defined:

1) **active - reflective**: How do you process information: actively (through physical activities and discussions), or reflexively (through introspection)?
2) **sensing-intuitive**: What kind of information do you tend to receive: sensitive (external agents like places, sounds, physical sensation) or intuitive (internal agents like possibilities, ideas, through hunches)?
3) **visual-verbal**: Through which sensorial channels do you tend to receive information more effectively: visual (images, diagrams, graphics), or verbal (spoken words, sound)?
4) **sequential - global**: How do you make progress: sequentially (with continuous steps), or globally (through leaps and an integral approach)?

The idea in the 1988 article was that many students have a visual way of learning, and then teachers should spend time devising visual aids (in addition to verbal aids - that were the prominent aids in 1988), and so on.

However, studies show that the students should use *many* different learning resources - not only one favourite (not only go to plenary lectures or not only read in the book).

Active Learning

Since active students are more able to analyse, evaluate and synthesise ideas

- Provide learning environments, opportunities, interactions, tasks and instruction that foster deep learning.
- Provide guidance and support that challenges students based on their current ability.
- Students discover their current strengths and weaknesses and what they need to do to improve.

What are student active learning methods/tasks?

- Pause in plenary lecture to ask questions and let students think and/or discuss.
- In-class quizzes (with the NTNU invention Kahoot!) — individual and team mode.
- Projects — individual or in groups.
- Group discussion.

Now: plenary and *interactive lectures*.

Learning resources in the GLM course

Different learning resources in this GLM course have been designed, hopefully many of these match your way of learning.

The module pages

The course is divided into modular units with specific focus, in order to use smaller units (time and topic) to facilitate learning.

- The topic of each module on the agenda for 1—2 weeks of study.
- All activity points to module pages.
- Mathematics in LaTeX (also derivations present), figures and examples with R, all R code visible.

Structure of module pages

1) Introduction and aim
2) Motivating example
3) Theory—example loop
4) Recommended exercises
5) References, packages to install.

How to use the module pages

- A slides version (output: beamer_presentation) of the pages used in the plenary lectures.
- A webpage version (output: html_document) used in the (so-called) interactive lectures.
- A document version (output: pdf_document) used for student self study.
- The Rmd version — used as notebook to investigate changes to the R code.
- Additional class notes (written in class) linked in.

The module pages are the backbone of the course!

The plenary lectures (PL)

- for each module we start with a plenary lecture to introduce the aims,
- use real data to exemplify what to learn, why this is useful and what this is used in society
- theory is then presented (writing - not slides), discussed and
- mixed with use of R and data analysis.

The plenary lectures is rather passive in nature - for the students - and held in classical auditorium. They provide the first step into the new module.

Questions

What are advantages of attending a plenary lecture (as compared to reading the text book or the module pages, or watching videos)? Do you plan to attend the plenary lectures?

The interactive lectures (IL)

Has focus on student activity and understanding though discussing with fellow students and with the lecturer/TA - in groups.

1. Students arrive and are divided into groups (different criteria will be used). Short presentation round (name, study programme, interests) in the groups. One student (the "manager") log in to the PC at each table, or connect her/his own laptop and display the module page.
2. Lecturer gives a *short* introduction to current state, and present a problem set (mainly exam problem).

3. Students work together in the group on the problem set. The problems are presented on the digital screen, and the students discuss by interacting around the screen and often by running (ready-made) R code and interpreting analysis output - all presented on the digital screen.

4. If the problem is of a theoretical flavour, or drawing is needed - the students work on the whiteboards next to the digital screen. One student may act as “secretary”.

5. Lecturer summarizes solutions to the problem with input from the student groups.
6. This summarizing the first 45 minutes, then there is a break and then repeat 1-5 in the second hour.

Questions:

- Who are the interactive lectures for?
- What are advantages of attending an interactive lecture?
- When you finish your studies and head for a job - go you think the skills developed in the interactive lectures will be in demand?
- Do you think the interactive lectures will be challenging for you to attend? Why?
- How can the lecturer help you make this easier? Personal adjustment can be made.

The compulsory exercises

Has mainly focus on programming and interpretation - with some theory - and can be worked on in small groups (1-3). Will be a test of acquired understanding, and will constitute 30% of the final evaluation.

Practical details

go to Blackboard student log-in or guest access.

End of Presentation Part

Core concept: Exponential family of distributions

Now lets's start with GLMs...

In this course we will look at models where the distribution of the response variable, y_i, can be written in the form of a *univariate exponential family*

$$f(y_i \mid \theta_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} \cdot w_i + c(y_i, \phi, w_i)\right)$$

where

- θ_i is called the canonical parameter and is a parameter of interest
- ϕ is called a nuisance parameter (and is not of interest to us=therefore a nuisance (plage))
- w_i is a weight function, in most cases $w_i = 1$
- b and c are known functions.

It can be shown that $\mathrm{E}(Y_i) = b'(\theta_i)$ and $\mathrm{Var}(Y_i) = b''(\theta_i) \cdot \frac{\phi}{w}$.

Remark: slightly different versions of writing the exponential family exists, but we will use this version in our course (a different version might be used in TMA4295, but the basic findings are the same).

Interactive lectures - problem set

You may of cause read through the problem set before the interactive lecture, but that is not a prerequisite. Solutions will be provided to the major part of the recommended exercises (but not to the R-part of this one).

Theoretical questions (first hour)

We will work with the exponential family, but to make the notation easier for these tasks, we omit the i subscript.

$$f(y \mid \theta) = \exp\left(\frac{y\theta - b(\theta)}{\phi} \cdot w + c(y, \phi, w)\right)$$

Problem 1:

Choose (discuss and then talk to lecturer/TA) if you will work on a) binomial, b) Poisson, c) univariate normal or d) gamma.

a) What process can produce a Y that is binomially distributed? Write down the probability mass function, f(x). Is the binomial distribution an the exponential family? Identify b and c and show the connection with the mean and variance of Y.

NB: you may first use $n = 1$ in the binomial (which then is called Bernoulli) - since that is much easier than a general n.

Hint: https://wiki.math.ntnu.no/tma4245/tema/begreper/discrete and nearly the same parameterization for showing the binomial is member of exponential https://www.youtube.com/watch?v=7mNrsFr7P_A.

b) What about the Poisson distribution? What process can produce a Y that is Poisson distributed? Write down the probability mass function, $f(x)$. Is the Poisson distribution an the exponential family? Identify b and c and show the connection with the mean and variance of Y.

Hint: https://wiki.math.ntnu.no/tma4245/tema/begreper/discrete and first part of Sannsynlighetsmaksimering

c) What about the (univariate) normal? What process can produce a Y that is normally distributed? Write down the probability distribution function, $f(x)$. Is the univariate normal distribution an the exponential family? Identify b and c and show the connection with the mean and variance of Y.

d) What about the gamma distribution? What process can produce a Y that is gamma distributed? There are many different parameterizations for the gamma pdf, and we will use this (our textbook page 643): $Y \sim Ga(\mu, \nu)$ with density

$$f(y) = \frac{1}{\Gamma(\nu)}(\frac{\nu}{\mu})^{\nu} y^{\nu-1} \exp(-\frac{\nu}{\mu}y) \text{ for } y > 0$$

Is the gammadistribution an the exponential family? Identify b and c and show the connection with the mean and variance of Y.

Hint:
https://wiki.math.ntnu.no/tma4245/tema/begreper/continuous

Problem 2. Choose either alternative a or b.

Alternative a: Prove that $\mathrm{E}(Y_i) = b'(\theta_i)$ and $\mathrm{Var}(Y_i) = b''(\theta_i) \cdot \frac{\phi}{w}$.

Hint: integration by parts, and investigate what is $\int_{-\infty}^{\infty} \frac{df}{dy} dy$?

Alternative b: The following is a derivation of the mean and variance of an exponential family. Go through this derivation and specify why you go from one step to another.

Derivation

Exam questions with the exponential family – optional (covered above)

We have covered the Poisson and gamma in the problem sets above, but not the negative binomial (not in the core of the course)

Exam December 2017, Problem 1a: Poisson regression

(Remark: last question can not be answered before module 4.)
Consider a random variable Y. In our course we have considered the univariate exponential family having distribution (probability density function for continuous variables and probability mass function for discrete variables)

$$f(y) = \exp(\frac{y\theta + b(\theta)}{\phi} w + c(y, \phi, w))$$

where θ is called the *natural parameter* (or parameter of interest) and ϕ the *dispersion parameter*.
The Poisson distribution is a discrete distribution with probability mass function

$$f(y) = \frac{\lambda^y}{y!} \exp(-\lambda), \text{ for } y = 0, 1, ...,$$

where $\lambda > 0$.

a) [10 points]

Show that the Poisson distribution is a univariate exponential family, and specify what are the elements of the exponential family $(\theta, \phi, b(\theta), w, c(y, \phi, w))$.

What is the connection between $\mathrm{E}(Y)$ and the elements of the exponential family?

What is the connection between $\mathrm{Var}(Y)$ and the elements of the exponential family?

Use these connections to derive the mean and variance for the Poisson distribution.

If the Poisson distribution is used as the distribution for the response in a generalized linear model, what is then the *canonical link* function?

Exam 2012, Problem 3: Precipitation in Trondheim, amount

Remark: the text is slightly modified from the original exam since we parameterized the gamma as in our textbook.

We want to model the amount of daily precipitation given that it *is* precipitation, and denote this quantity Y. It is common to model Y as a gamma distributed random variable, $Y \sim Gamma(\nu, \mu)$, with density

$$f_Y(y) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1} \exp\left(-\frac{\nu}{\mu} y\right)$$

In this problem we consider N observations, each gamma distributed with $Y_i \sim Gamma(\nu, \mu_i)$ (remark: common ν). Here ν is considered to be a known nuisance parameter, and the μ_is are unknown.

a) Show that the gamma distribution function is member of the exponential family when μ_i is the parameter of interest.
Use this to find expressions for the expected value and the variance of Y_i, in terms of (ν, μ_i), and interpret ν.

Exam 2010, Problem 2: Negative binomial distribution

The probability density function for a negative binomial random variable is

$$f_y(y;\theta,r) = \frac{\Gamma(y+r)}{y!\Gamma(r)}(1-\theta)^r\theta^y$$

for $y = 0, 1, 2, ...,$, $r > 0$ and $\theta \in (0, 1)$, and where $\Gamma()$ denotes the gamma function. (There are also other parameterizations of the negative binomial distributions, but use this for now.)

a) Show that the negative binomial distribution is an exponential family. You can in this question consider r as a known constant.

b) Use the general formulas for a exponential family to show that $\mathsf{E}(Y) = \mu = r\frac{\theta}{1-\theta}$ and $\mathsf{Var}(Y) = \mu\frac{1}{1-\theta}$.

Focus on R-related topics (second hour)

R, Rstudio, CRAN and GitHub - and R Markdown

What is R?

https://www.r-project.org/about.html

What is Rstudio?

https://www.rstudio.com/products/rstudio/

What is an R package?

http://r-pkgs.had.co.nz (We will make an R package in the exercise part of this course.)

What is CRAN?

https://cran.uib.no/

What is GitHub and Bitbucket?

Do we need GitHub or Bitbucket in our course?
https://www.youtube.com/watch?v=w3jLJU7DT5E and https://techcrunch.com/2012/07/14/what-exactly-is-github-anyway/

What is R Markdown?

http://r4ds.had.co.nz/r-markdown.html

What is `knitr`?

https://yihui.name/knitr/

What is R Shiny?

https://shiny.rstudio.com/
(In the statistics group we will build R Shiny app for the thematic pages for our TMA4240/TMA4245/ST1101/ST1201/ST0103 introductory courses, so if you have ideas for cool graphical presentation please let us know - we have some economical resources available for help from master students in statistics! Also ideas for this GLM course is or interest!)
The IMF R Shiny server is here: https://shiny.math.ntnu.no/ (not anything there now, but a lot more soooon).
(Remember the test you did to brush up on R programming? https://tutorials.shinyapps.io/04-Programming-Basics/#section-welcome This was made with a combination of the R package `learnr` and a shiny server.)

Explore R Markdown in Rstudio

Quotations from
https://rmarkdown.rstudio.com/authoring_quick_tour.html:

- Creating documents with R Markdown starts with an .Rmd file that contains a combination of markdown (content with simple text formatting) and R code chunks.
- The .Rmd file is fed to knitr, which executes all of the R code chunks and creates a new markdown (.md) document which includes the R code and it's output.
- The markdown file generated by knitr is then processed by pandoc which is responsible for creating a finished web page, PDF, MS Word document, slide show, handout, book, dashboard, package vignette or other format.

The module pages (you are reading the Module 1 page now), are written using R Markdown. To work with the module pages you either copy-paste snippets of R code from the module page over in your editor window in Rstudio, or copy the Rmd-version of the module page (1Intro.Rmd) into your Rstudio editor window (then you can edit directly in Rmarkdown document - to make it into your personal copy).

If you choose the latter: To compile the R code we use `knitr` (termed "knit") to produce a html-page you press "knit" in menu of the editor window, but first you need to install packages: `rmarkdown` and `devtools` (from CRAN). For the module pages the needed R packages will always be listed in the end of the module pages.

If you want to learn more about the R Markdown (that you may use for the compulsory exercises) this is a good read:

- http://r4ds.had.co.nz/r-markdown.html (Chapter 27: R Markdown from the "R for Data Science" book), and
- the Rstudio cheat sheet on R Markdown is here:

Not use R Markdown, but only R code?

If you only want to extract the R code from a R Markdown file you may do that using the function `purl` from library `knitr`. To produce a file "1Intro.R" from this "1Intro.Rmd" file:

```
library(knitr)
purl("https://www.math.ntnu.no/emner/TMA4315/2018h/1Intro.R
```

The file will then be saved in your working directory, that you see with `getwd()`.

R packages

And to work with either the 1Intro.R or 1Intro.Rmd file you will have to first install the following libraries:

```
install.packages(c("rmarkdown","gamlss.data","tidyverse","g
```

For the subsequent module pages this information will be available in the end of the page.

The Munich Rent Index Data set

We will use this data set when working with multiple linear regression (next module), so this is a good way to start to know the data set and the ggplot functions, which can be installed together with a suite of useful libraries from `tidyverse`.
A version of the Munich Rent Index data is available as `rent` in library `catdata` from CRAN.

```
library(gamlss.data)
library(ggplot2)
```

Get to know the `rent` data.

```
ds=rent99
colnames(ds)
```

```
## [1] "rent"     "rentsqm"  "area"     "yearc"    "locatio
## [8] "cheating" "district"
```

```
dim(ds)
```

```
## [1] 3082    9
```

```
summary(ds)
```

Combining exercise 1 and 2:

Choose one of the distributions you studied earlier (binomial, Poisson, normal or gamma), and write a R-markdown document answering the questions on requirements, f(x), f(x) as exponential family and mean and variance. Also add R-code to plot f(x) and F(x) for a given set of parameters - and add the mean as a vertical line - using the ggplot library. Submit your Rmd document to the lecturer (email) - so it can be added to this module solutions, or make your own github repository and email the link to your repo to be added to this module page.

Further reading

- Grolemund and Hadwick (2017): "R for Data Science", http://r4ds.had.co.nz
- Xie, Allaire and Grolemund (2018): "R Markdown — the definitive guide", https://bookdown.org/yihui/rmarkdown/
- Hadwick (2009): "ggplot2: Elegant graphics for data analysis" textbook.
- Wilkinson (2005): The grammar of graphics. The theory behind the ggplot2 package universe.
- If you want to see more of the powers of ggplot, combined with a nice story: https://www.andrewheiss.com/blog/2017/08/10/exploring-minards-1812-plot-with-ggplot2/
- R-bloggers: https://https://www.r-bloggers.com/ is a good place to look for tutorials.
- Stack Overflow: https://stackoverflow.com/ is a good place