Module 1: INTRODUCTION

TMA4315 Generalized linear models H2018

Mette Langaas, Department of Mathematical Sciences, NTNU

23.08 [PL] and 24.08 [IL]

Contents

Introduction

Classnotes 23.08.2018

Aim of this module

- this course: expanding the linear regression framework
- short presentation of all course modules
- learning outcome
- student learning styles
- interactive lectures: what, why and how?
- practical details of the course (Blackboard)
- 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

---

Expanding the linear regression framework

(see classnotes)

You know multiple linear regression (from TMA4267 Linear statistical models or TMA4255 Applied Statistics or TMA4268 Statistical learning). 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.

We will consider relationships between the *conditional mean of* Y_i, $\mathrm{E}(Y_i \mid \mathbf{x}_i) = \mu_i$, and linear combinations of the covariates in *a linear predictor* $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} = \mathbf{x}_i^T \beta$.

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).

Course content and modules

Univariate exponential family. Multiple linear regression. Logistic regression. Poisson regression. General formulation for generalised linear models with canonical link. Likelihood-based inference with score function

and expected Fisher information. Deviance. AIC. Wald and likelihood-ratio test. Linear mixed effects models with random components of general structure. Random intercept and random slope. Generalised linear mixed effects models. Strong emphasis on programming in R.

Possible extensions: quasi-likelihood, over-dispersion, models for multinomial data, analysis of contingency tables, quantile regression.

H2018 extensions: categorical regression (models for multinomial data) and contingency tables, score tests.

The modules - in short

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.

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]

Module 2: Multiple linear regression

Example: Exam TMA4267, V2017, Problem 2: CVD

The Framingham Heart Study is a study of the etiology (i.e. underlying causes) of cardiovascular disease (CVD), with participants from the community of Framingham in Massachusetts, USA https://www.framinghamheartstudy.org/. This dataset is subset of a teaching version of the Framingham data, used with permission from the Framingham Heart Study.

We will focus on modelling systolic blood pressure using data from $n = 2600$ persons. For each person in the data set we have measurements of the following seven variables

- `SYSBP` systolic blood pressure (mmHg),
- `SEX` 1=male, 2=female,
- `AGE` age (years) at examination,
- `CURSMOKE` current cigarette smoking at examination: 0=not current smoker, 1= current smoker,

- `BMI` body mass index (kg/m^2),
- `TOTCHOL` serum total cholesterol (mg/dl), and
- `BPMEDS` use of anti-hypertensive medication at examination: 0=not currently using, 1=currently using.

A multiple normal linear regression model was fitted to the data set with $-\frac{1}{\sqrt{SYSBP}}$ as response and all the other variables as covariates.

The data set is here called `thisds`.

```
modelB=lm(-1/sqrt(SYSBP)~SEX+AGE+CURSMOKE+BMI+TOTCHOL+BPMEDS,data=thisds)
summary(modelB)
```

```
##
## Call:
## lm(formula = -1/sqrt(SYSBP) ~ SEX + AGE + CURSMOKE + BMI + TOTCHOL +
##     BPMEDS, data = thisds)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0207366 -0.0039157 -0.0000304  0.0038293  0.0189747
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.103e-01  1.383e-03 -79.745  < 2e-16 ***
## SEX         -2.989e-04  2.390e-04  -1.251 0.211176
## AGE          2.378e-04  1.434e-05  16.586  < 2e-16 ***
## CURSMOKE    -2.504e-04  2.527e-04  -0.991 0.321723
## BMI          3.087e-04  2.955e-05  10.447  < 2e-16 ***
## TOTCHOL      9.288e-06  2.602e-06   3.569 0.000365 ***
## BPMEDS       5.469e-03  3.265e-04  16.748  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.005819 on 2593 degrees of freedom
## Multiple R-squared:  0.2494, Adjusted R-squared:  0.2476
## F-statistic: 143.6 on 6 and 2593 DF,  p-value: < 2.2e-16
```

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 respons that is not a continuous variable? Here we look at present/absent, true/false, healthy/diseased.

Example: Mortality of beetles

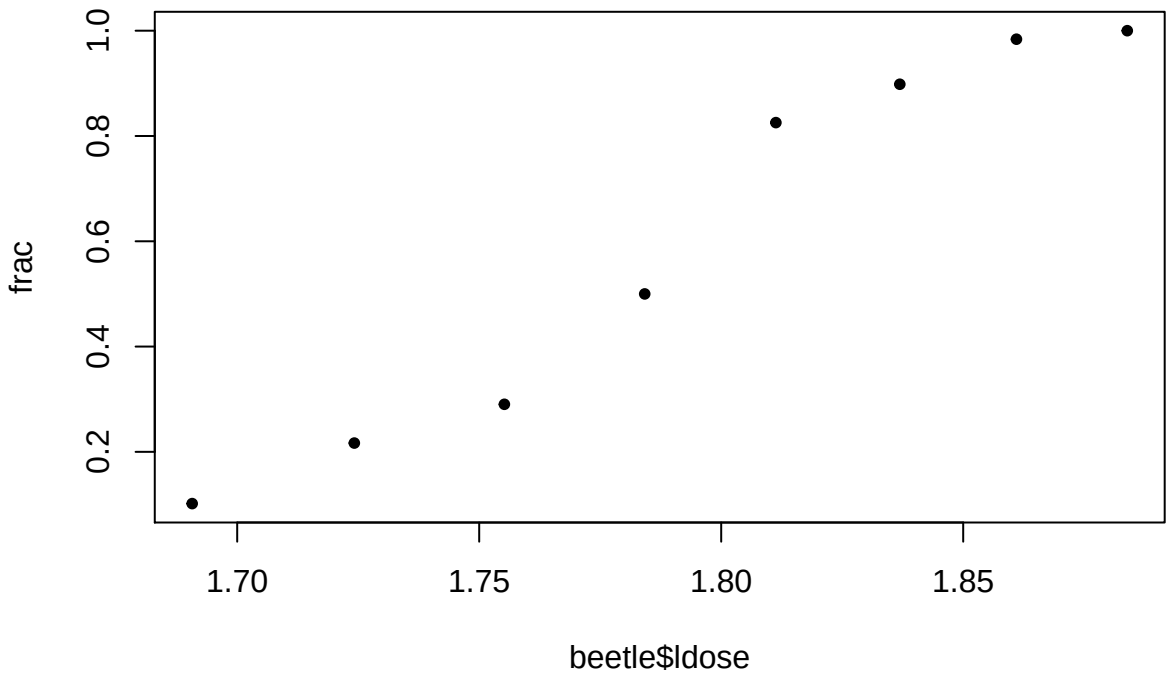
About 60 beetles were exposed to each of 8 different concentrations of CS_2 (data on log10-dose), and the number killed at each of the concentrations were recorded.

```
library(investr)
head(beetle)
```

```
##    ldose  n  y
## 1 1.6907 59  6
## 2 1.7242 60 13
## 3 1.7552 62 18
## 4 1.7842 56 28
## 5 1.8113 63 52
```

```
## 6 1.8369 59 53
```

```
frac=beetle$y/beetle$n
plot(beetle$ldose,frac,pch=20)
```



What might be the distribution of the number of dead beetles, Y_i at a given dose x_i? Dose x_i was given to n_i beetles.

$$Y_i = bin(n_i, \pi_i)$$

where π_i =probability for a beetle to die at dose x_i and n_i= number of beetles treated with dose x_i. A linear model for π_i estimated by ordinary least squares is problematic because

- $0 \leq \pi_i \leq 1$ can not be guaranteed by a linear expression $\beta_0 + \beta_1 x_i$, and
- $\text{Var}(Y_i) = n_i \pi_i (1 - \pi_i)$ is non-constant (heteroscedastic) variance.

The "usual" solution to this is *logistic regression* where the relationship between the mean of the response and the predictor is not linear, but instead

$$\ln(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 x_i$$

or equivalently

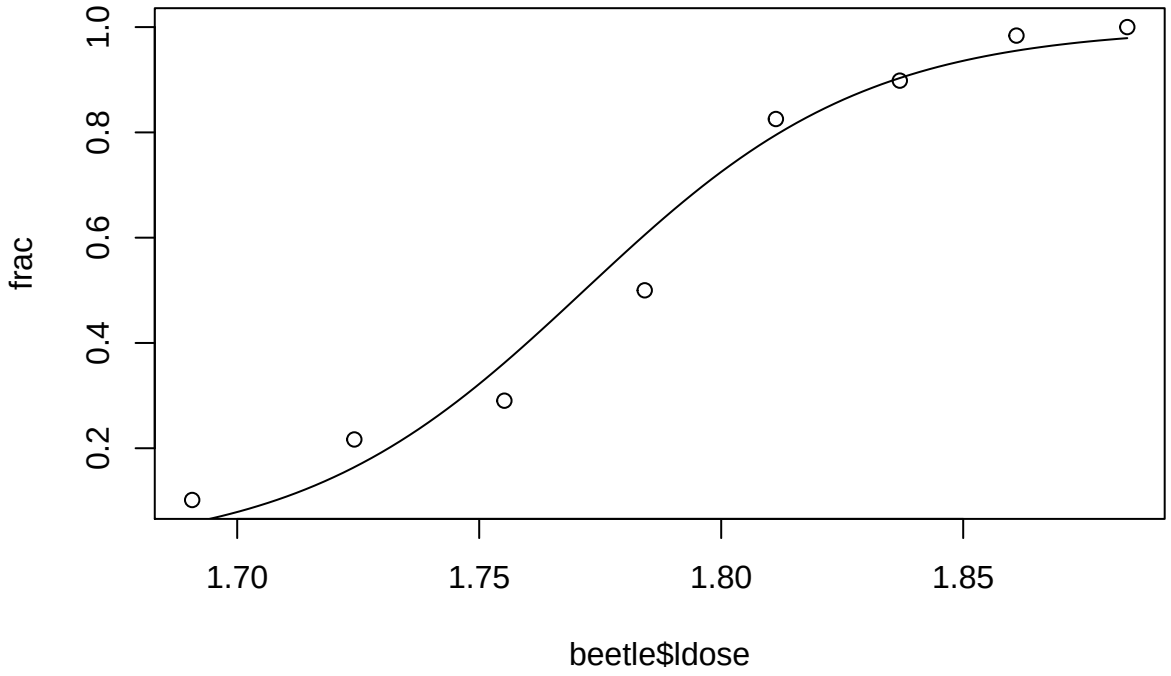
$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

Then $0 \leq \pi_i \leq 1$. We estimate the model by Maximum Likelihood (ML), while taking into account that the responses are binomially distributed.

```
fit=glm(cbind(beetle$y,beetle$n-beetle$y)~ldose,data=beetle,family=binomial)
summary(fit)
```

```
##
## Call:
## glm(formula = cbind(beetle$y, beetle$n - beetle$y) ~ ldose, family = binomial,
##     data = beetle)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5941  -0.3944   0.8329   1.2592   1.5940
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -60.717      5.181  -11.72   <2e-16 ***
## ldose         34.270      2.912   11.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 284.202  on 7  degrees of freedom
## Residual deviance:  11.232  on 6  degrees of freedom
## AIC: 41.43
##
## Number of Fisher Scoring iterations: 4
```

```
thisrange=range(beetle$ldose)
xs=seq(thisrange[1],thisrange[2],length=100)
predicted=predict(fit,newdata=data.frame(ldose=xs),type="response")
plot(beetle$ldose,frac)
lines(xs,predicted)
```



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. In one famous study British doctors were in 1951 sent a questionnaire about whether they smoked tobacco - and later information about their death were collected. Questions that were asked were: Is the death rate higher for smokers than for non-smokers? If so, by how much? And, how is this related to age?

```r
library(boot) #n=person-year, ns=smoker-years, age=midpoint 10 year age group,
#y=number of deaths due to cad, smoke=smoking status
head(breslow,n=10)
```

```
##    age smoke     n   y    ns
## 1   40     0 18790   2     0
## 2   50     0 10673  12     0
## 3   60     0  5710  28     0
## 4   70     0  2585  28     0
## 5   80     0  1462  31     0
## 6   40     1 52407  32 52407
## 7   50     1 43248 104 43248
## 8   60     1 28612 206 28612
## 9   70     1 12663 186 12663
## 10  80     1  5317 102  5317
```

To investigate this we will look at different ways of relating the expected number of deaths and the number of doctors at risk in the observation period for each smoke and age group. We will do this by assuming a Poission distribution for the number of deaths, and linking this to a linear predictor.

When we work with continuous data - like life times, costs and claim sized - these may not be negative, and the their distribution often follow a right skewed distribution. We will look at effect a one of more covariates that may work multiplicative on the response and see how we may fit that using gamma regression 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:

1. The mean of the response, $\mu_i = \mathrm{E}(Y_i)$, is connected to the linear predictor $\eta_i = \mathbf{x}_i^T\beta$ by a link function: $\eta_i = g(\mu_i)$ or, alternatively, by a response function $\mu_i = h(\eta_i)$ - where $g = h^{-1}$ (inverse functions).
2. The distribution of the response can be written as a univariate exponential family (we work with that in this first module).

This leads to a unified framework, and maximum likelihood estimation can be written on a generalized form for all the GLMs. In addition we can present 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) and solve using generalized estimation equations.

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. Examples of categorical responses include (unordered) data in infection (no, or socalled type I or type II) after Caesarian delivery, or (ordered) data on degree of defoliation of trees (nine ordered categories).

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.

```r
ds=read.table("https://www.math.ntnu.no/emner/TMA4315/2017h/data/caesarian.raw",header=TRUE)
head(ds)
```

```
##    n infbin RISK NPLAN ANTIB Y
## 1  0      1    1     0     1 1
## 2  1      1    1     0     1 2
## 3 17      0    1     0     1 3
## 4  0      1    0     0     1 1
## 5  0      1    0     0     1 2
## 6  2      0    0     0     1 3
```

For unordered categories (like the Caesarian delivery data) we will use many logistic regressions - each between one category and a chosen reference category. For ordered categories (like the defoliation of trees) we will use a cumulative model, also called a proportional odds model.

If time permits we will also look briefly at exact and asymptotic inference (Fishers exact test and Pearsons Chisquare test) for contingency tables (unordered categories), which is closely related to the GLM-presentation.

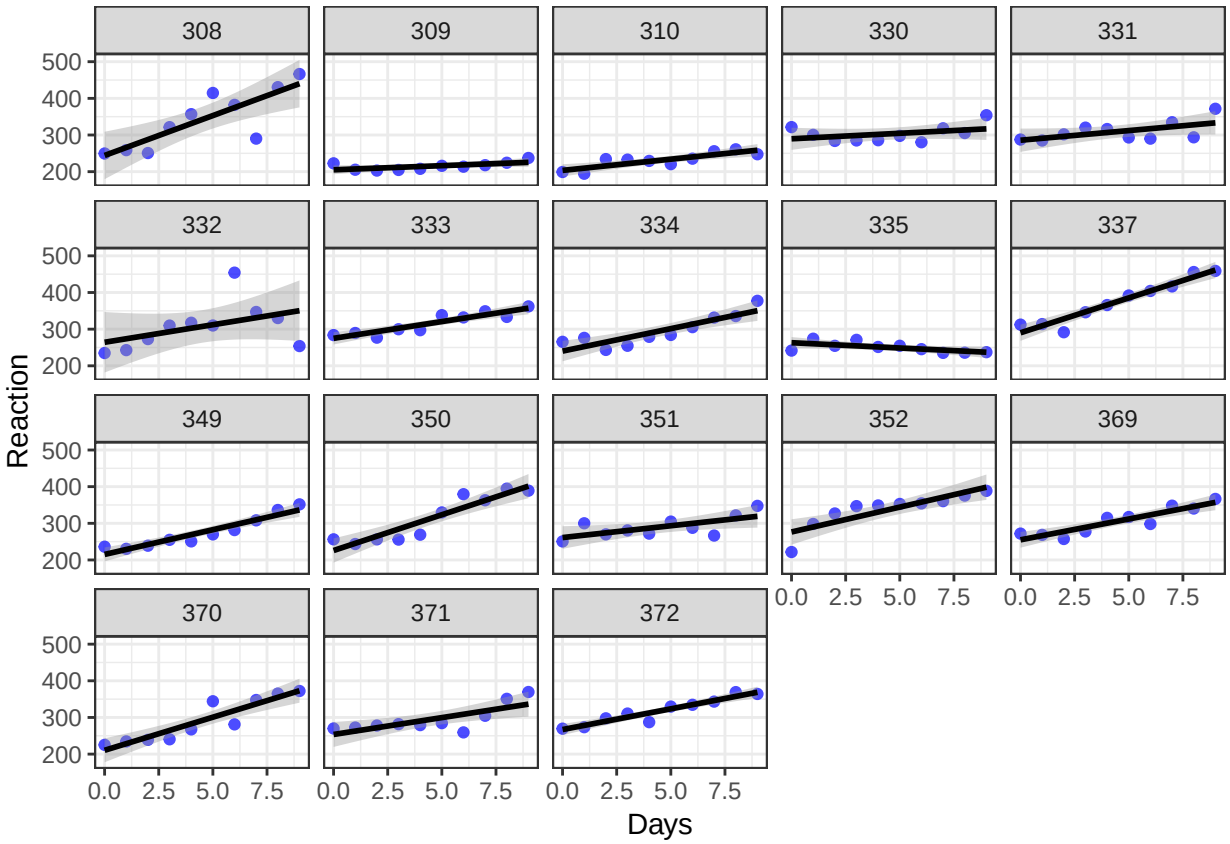
Textbook: Chapter 6, and possibly extra materiale on the Fisher and Chi-square test (if time permits).

Compulsory exercise 2 will cover modules 3-6.

Module 7: Linear mixed effects models

In a study on the effect of sleep deprivation the average reaction time per day were measured. On day 0 the subjects had their normal amount of sleep. Starting that night they were restricted to 3 hours of sleep per night. The observations represent the average reaction time on a series of tests given each day to each subject. This was measured for 18 subjects for 10 days (days 0-9).

```
library(lme4)
library(ggplot2) # see more on ggplot later in this module
gg <- ggplot(sleepstudy, aes(x = Days, y = Reaction))
gg <- gg + geom_point(color = "blue", alpha = 0.7)
gg <- gg + geom_smooth(method = "lm", color = "black")
gg <- gg + theme_bw()
gg <- gg + facet_wrap(~Subject)
gg
```



We observe that each subject's reaction time increases approximately linearly with the number of sleepdeprived days. But, it appears that subjects have different slopes and intercepts.

As a first model we may assume that there is a common intercept and slope for the population - called fixed effects, but allow for random deviations for the intercept and slope for each individual. In linear mixed effects models we assume that the random intercepts and slopes are drawn from normal distributions and estimate the variance in these distribution. Such a model will make observations correlated within subjects.

```
fm1 <- lmer(Reaction ~ Days + (Days | Subject), sleepstudy)
summary(fm1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Reaction ~ Days + (Days | Subject)
##    Data: sleepstudy
##
## REML criterion at convergence: 1743.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.9536 -0.4634  0.0231  0.4634  5.1793
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  Subject  (Intercept) 612.09   24.740
##           Days         35.07    5.922   0.07
##  Residual             654.94   25.592
## Number of obs: 180, groups:  Subject, 18
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  251.405      6.825  36.838
## Days          10.467      1.546   6.771
##
## Correlation of Fixed Effects:
##      (Intr)
## Days -0.138
```

Here the population fixed effects estimates are an intercept of 251.4 ms and a slope of 10.47 ms/day. The random effects for the intercept and the slope have estimated standard deviations 24.74 ms and 5.92 ms/day.

In this module we will look at different models for clustered (schools, families) 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

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.

Learning outcome

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 generelized 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.

Learning styles

We (probably) all have different ways in which we learn - and we have different learning ambitions when attending a course.

Back in 1988 Felder and Silverman published an article where they suggested that there was a mismatch between the way students learn and the way university courses were taught (in Science, Technology, Engineering and Mathematics=STEM). They 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)?

Here are a few words on the four axis

The idea in the 1988 article was that by *acknowledging these different learning style axes it was possible to guide the teachers to choose teaching styles that matched the learning styles of the students.* That is, many students (according to Felder and coauthors) 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).

In this GLM course I have designed different learning resources, and hope that many of these match your way of learning. To help me (and maybe for you to get some insight into your own learning style)

I ask you to answer a standardized set of questions made by Felder et al (44 questions with two possible answers), and then report your results to me in a Google form.

You can report your scores anonymously, or by giving our name. If you do this anonymously I will have information on the class level, and if you do this with your name I get to know a bit about how you learn and I can use the results to help to construct student groups for the interactive lectures. Your results will only be used by me, and I will not show them to other people (students or staff). This means that this is not used for research, but to increase the quality of the GLM course (in total and for each one of you). I will never discuss your personal results in class, but are very eager to discuss results on the class level - and use these when designing new learning resources.

Here is the questionarie (maybe do a screen shot of your results -the results only appear on a web page and is not saved or emailed to you or anyone).

I have taken the test and these were my results: I scored 3 on the active side of the active-reflective scale, 1 on the sensing side of the sensing-intuitive scale, 5 on the visual side on the visual-verbal scale and finally 5 on the global side of the sequential-global scale. In the Google form I would then report to have a “active value” for the active-reflective axis, and then report the value to be 3. Then I would choose “sensing value” on the sensoring-intuitive axis and report the value to be 1, I will choose “visual value” and report 5, and

finally choose "global value" and report 5. (Values below 5 are reported to be weak, and this means that there is no strong preference on that axis.)

Here is the Google form where I ask that you write your 4 scores

After you have submitted your scores please go back and read the description of the four axis, but this time focus on the advice for the different type of learners

If you are curios about the work of Felder and coauthors, more resources can be found here: http://educationdesignsinc.com/

Learning resources in the GLM course

The module pages

I have divided the GLM course 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!

Active students — deep 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 learing methods/tasks?

- Pause in plenary lecture to ask questions and let students think and/or discuss.

Figure 1:

- In-class quizzes (with the NTNU invention Kahoot!) — individual and team mode.
- Projects — individual or in groups.
- Group discussion.

Now: plenary and *interactive lectures*.

The plenary lectures (PL)

- for each module we start with a plenary lecture to introduce the aims,
- use real data to examplify what to learn, why this is useful and what this is used in society
- then we move to notation and focus on the model used
- 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.

Q: What are advantages of attending a plenary lecture (as compared to reading the text book or the module pages, or watching videoes)? 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.

Smia (the smithy)

A room where interaction and activity is in focus. Flat floor with group tables, whiteboard and screen — PC and electricity outlets. 50 students.

https://www.ntnu.no/laeringsarealer/smia

---

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 (with light refreshments) and then repeat 1-5 in the second hour.

More pictures of how the students in H2017 worked will be shown in the PL.

Statements from focus groups H2017

(two groups of 4 students each, one hour "interviews" by external evaluator - anonymous contributtions)

The concept

Student A: We are taught in a different way than what we have experienced earlier — sitting in groups, working on problem sets and discussing. No, never experienced this before — and we are master students. Where was all this earlier?

Student B: It is very nice to have two hours every week to discuss with the others, and be able to explain to each other and work with the course material from a new point of view. We are not told what is right, but we spend time finding that out — together.

They also commented that the reading list was short and that they did not think they had to prepare much for the exam — they felt that they really understood and were up to date.

Learning outcome in group setting and new concepts

Student C: For most sessions I feel I learnt a lot. I remember the concept we work on has been talked about in the plenary lectures, and then I talk about the details with the others in the group and get to explain to the others — then I feel that I really know this concept. I do not really learn so much new stuff, but I learn what we have already gone through in the plenary lectures a lot better.

Student D: And, we get a confirmation that we have understood what we were taught in the plenary lecture. Yes, I know this concept - and so on - and I have not misunderstood - which may often happen. If I have misunderstood I get corrected here in the interactive lecture - this makes the learning more targeted (not so abstract).

Student E: Yes, I agree, what we learn becomes reinforced. Personally I find it hard to learn new concepts in a group setting.

Comment: hard to come to IL if not up-to-date on reading list (e.g. not read by yourself or attended PL).

But, they were also worried:

Student F: I believe Mette cannot go so deep in the plenary lecture - compared to when we had the double amount of plenary lectures. She plans for us to learn by ourselves, which I think is a good thing. I beleive that we have a greater gain from learning together, and from seeing each others problems, and we learn from each other.

Student G: I agree, I think it is more challenging for a lecturer to divide the course in interactive and plenary lectures than only using plenary lectures. The lecturer needs to teach in two different ways, but also to try to cover all material in effectively shorter time. Maybe this results in us loosing the depth understanding, maybe that is how it is, and that is sad.

This was the main motivation behind the module pages — having the full story written out, but choosing parts to present in the plenary lectures and parts in the interactive lectures.

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.
- If the IL worked well in 2017, does it mean that it will also work in 2018?

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.

Core concept: Exponential family of distributions

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 $\mathrm{Var}(Y_i) = b''(\theta_i) \cdot \frac{\phi}{w}$.

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 $y = 0, 1, 2, \ldots,$, $r > 0$ 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 $\mathrm{E}(Y) = \mu = r\frac{\theta}{1-\theta}$ and $\mathrm{Var}(Y) = \mu\frac{1}{1-\theta}$.

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: https://www.rstudio.com/wp-content/uploads/2016/03/rmarkdown-cheatsheet-2.0.pdf.

Then you see that you can make a pdf-file in addition to a html-file (for your reports you may choose either). To make the pdf-file you need latex to be installed on your machine.

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.Rmd")
```

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","prettydoc","gamlss.data","tidyverse","ggpubr","investr","lme4"))
```

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

The Munic 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 Munic 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"    "location" "bath"
## [7] "kitchen"  "cheating" "district"
```

```
dim(ds)
```

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

```
summary(ds)
```

```
##       rent           rentsqm            area            yearc
##  Min.   :  40.51   Min.   : 0.4158   Min.   : 20.00   Min.   :1918
##  1st Qu.: 322.03   1st Qu.: 5.2610   1st Qu.: 51.00   1st Qu.:1939
##  Median : 426.97   Median : 6.9802   Median : 65.00   Median :1959
##  Mean   : 459.44   Mean   : 7.1113   Mean   : 67.37   Mean   :1956
```

```
##  3rd Qu.: 559.36   3rd Qu.: 8.8408   3rd Qu.: 81.00   3rd Qu.:1972
##  Max.   :1843.38   Max.   :17.7216   Max.   :160.00   Max.   :1997
##  location bath     kitchen  cheating    district
##  1:1794   0:2891   0:2951   0: 321   Min.   : 113
##  2:1210   1: 191   1: 131   1:2761   1st Qu.: 561
##  3:  78                               Median :1025
##                                       Mean   :1170
##                                       3rd Qu.:1714
##                                       Max.   :2529
```

Then, head for plotting with `ggplot` but first take a quick look at the `ggplot2` library:

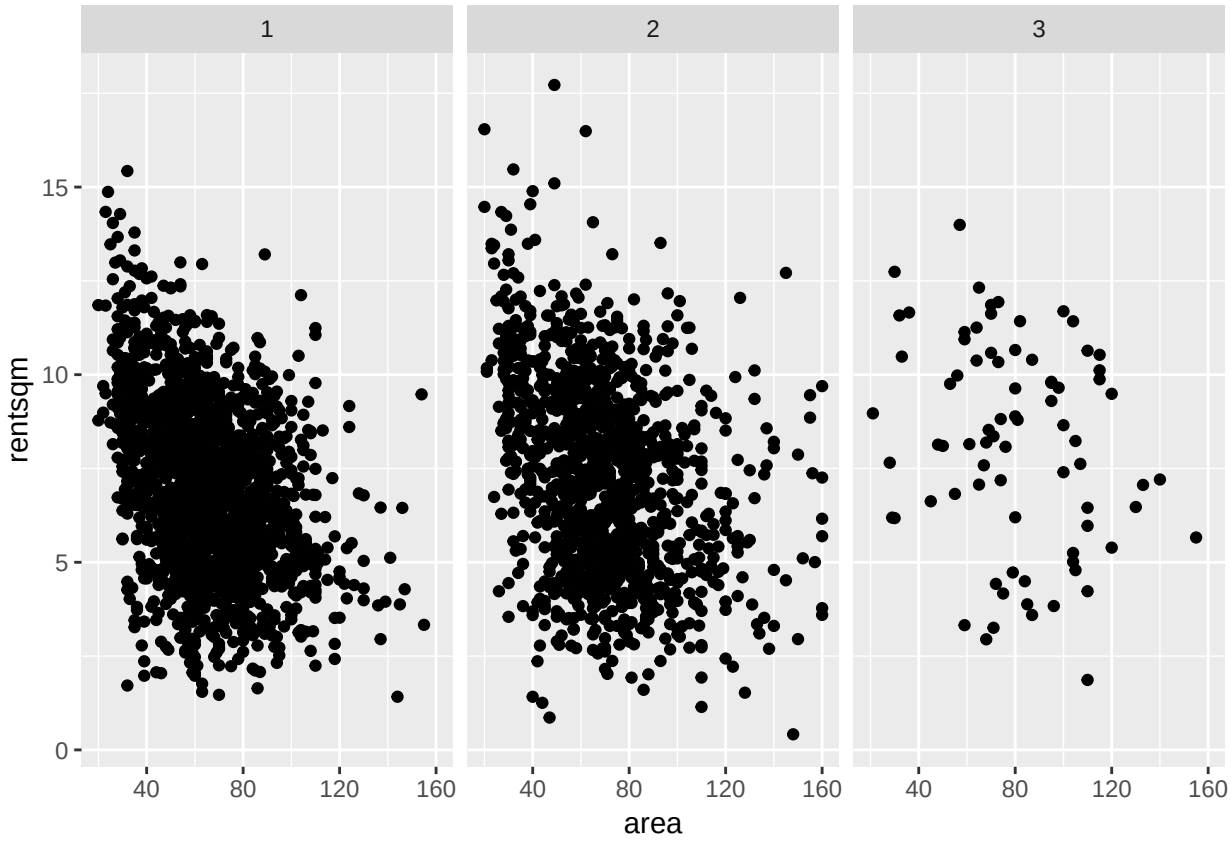
- Grolemund and Hadwick (2017): "R for Data Science", Chapter 3: Visualisation: http://r4ds.had.co.nz/data-visualisation.html

Before you continue you should have read the start of the Visualisation chapter that explains the ggplot grammar. Yes, you start with creating the coordinate system with `ggplot` and then add layers. What does the following words mean: mapping, aesthetic, geom function mean in the `ggplot` setting?

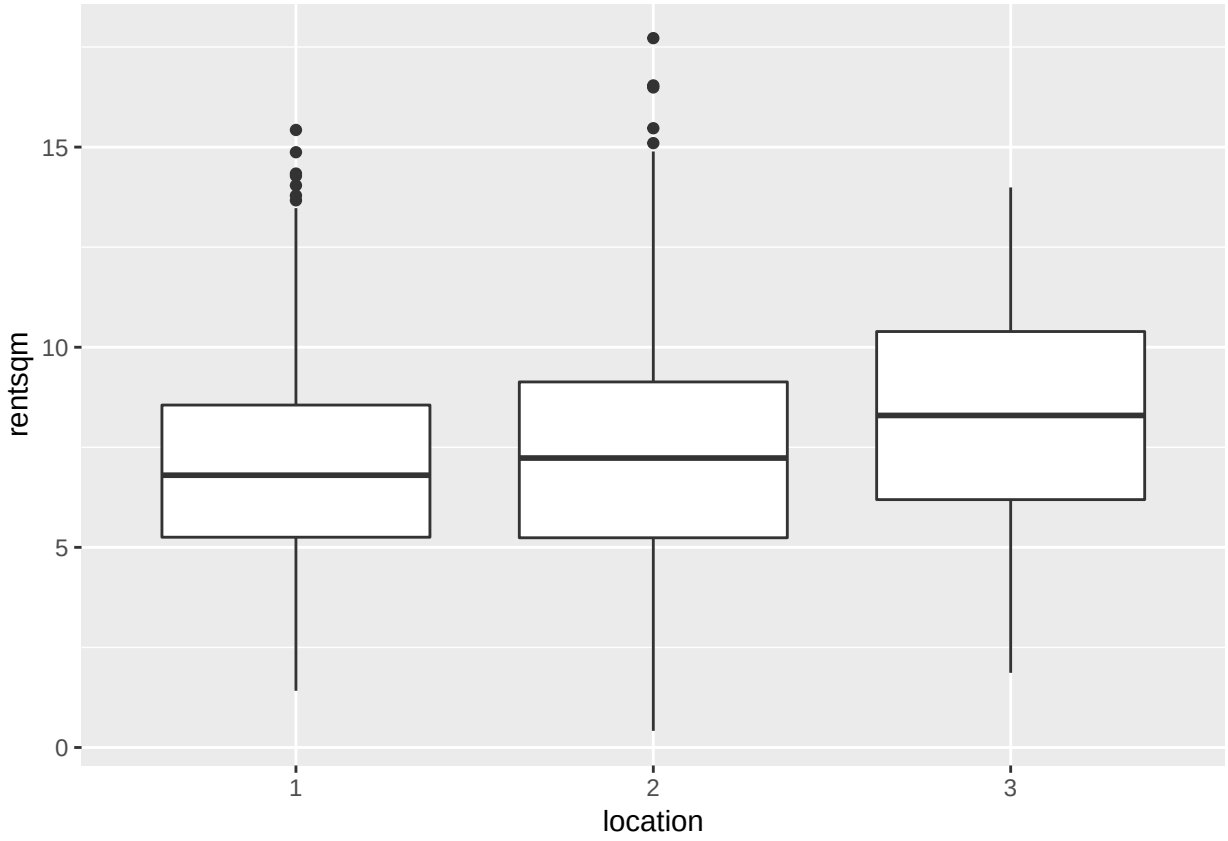
- The Rstudio cheat sheet on ggplot2 is here: https://www.rstudio.com/wp-content/uploads/2016/11/ggplot2-cheatsheet-2.1.pdf

First, look at plotting `rentsqm` for different values of `location` - with panels of scatter plots and with boxplots

```
ggplot(data=ds)+
  geom_point(mapping=aes(area,rentsqm))+
  facet_wrap(~location,nrow=1)
```



```
ggplot(data = ds, mapping = aes(x = location, y = rentsqm)) +
  geom_boxplot()
```



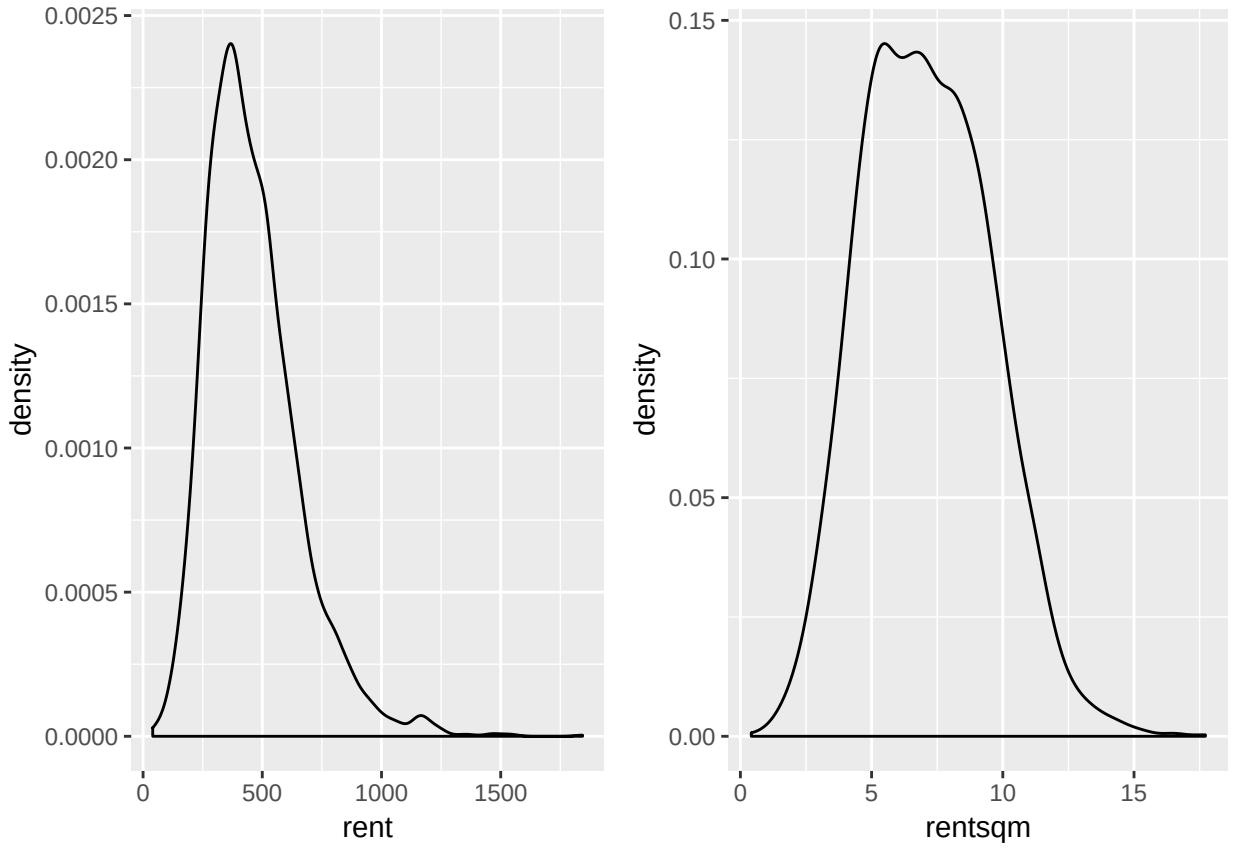
So, location matters.

But, should we use `rent` or `rentsqm` as response?

```
library(ggpubr)
```

```
## Loading required package: magrittr
```

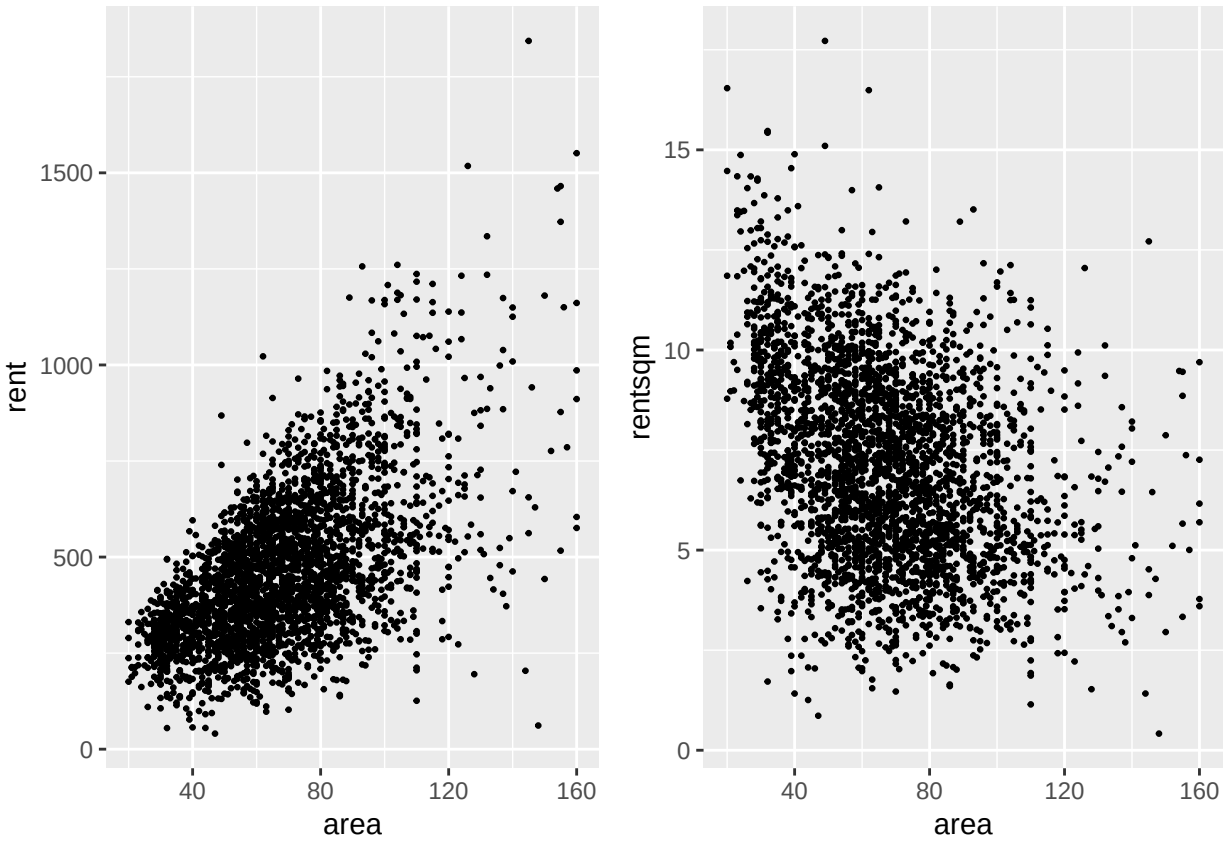
```
plot1 <- ggplot(data=ds) +
  geom_density(mapping=aes(rent),kernel="gaussian")
plot2 <- ggplot(data=ds) +
  geom_density(mapping=aes(rentsqm),kernel="gaussian")
ggarrange(plot1, plot2, ncol=2)
```



So, which response will we use? And, what if we would include `area` as covariate? I have plotted two plots together below, more on mixing graphs on the same page (we need ggprbr, gridExtra and cowplot packages) https://www.r-bloggers.com/ggplot2-easy-way-to-mix-multiple-graphs-on-the-same-page/

Relationship between `rent` or `rentsqm` and `area`

```
plot1 <- ggplot(data=ds,aes(area,rent))+
  geom_point(mapping=aes(area,rent),size=0.5)
plot2 <- ggplot(data=ds)+
  geom_point(mapping=aes(area,rentsqm),size=0.5)
ggarrange(plot1, plot2, ncol=2)
```



So, if we include area as a covariate, we may look at residuals when using `rent` or `rentsqm`. More about diagnostic plots in Module 2 - but - which plot below looks more random?

```
lm.rent=lm(rent~area,data=ds)
summary(lm.rent)
```

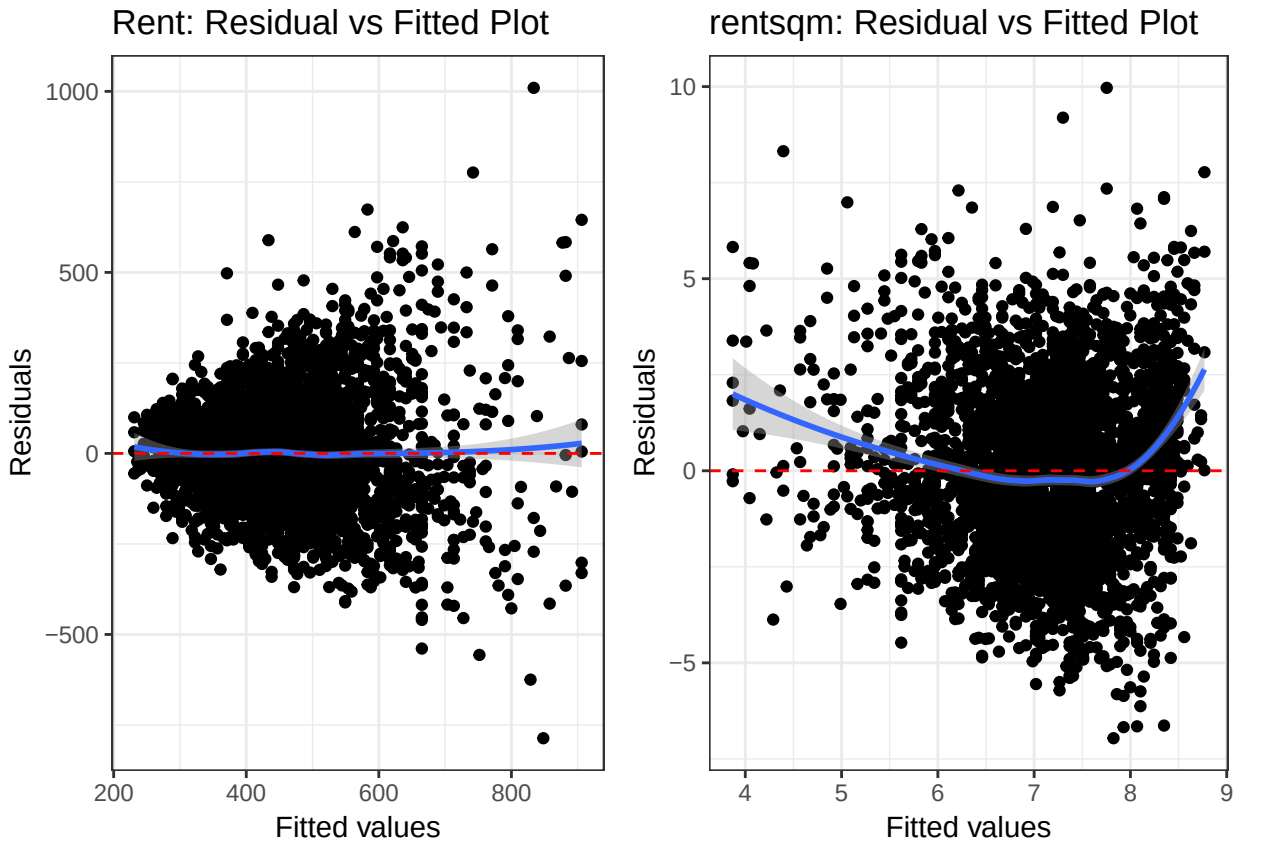
```
##
## Call:
## lm(formula = rent ~ area, data = ds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -786.63 -104.88   -5.69   95.93 1009.68
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 134.5922     8.6135   15.63   <2e-16 ***
## area          4.8215     0.1206   39.98   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 158.8 on 3080 degrees of freedom
## Multiple R-squared:  0.3417, Adjusted R-squared:  0.3415
## F-statistic:  1599 on 1 and 3080 DF,  p-value: < 2.2e-16
```

```
lm.rentsqm=lm(rentsqm~area,data=ds)
summary(lm.rentsqm)
```

```
##
```

```
## Call:
## lm(formula = rentsqm ~ area, data = ds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9622 -1.5737 -0.1102  1.5861  9.9674
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.46883    0.12426   76.20   <2e-16 ***
## area        -0.03499    0.00174  -20.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.291 on 3080 degrees of freedom
## Multiple R-squared:  0.1161, Adjusted R-squared:  0.1158
## F-statistic: 404.5 on 1 and 3080 DF,  p-value: < 2.2e-16
```

```
p1<-ggplot(lm.rent, aes(.fitted, .resid))+geom_point()
p1<-p1+stat_smooth(method="loess")+geom_hline(yintercept=0, col="red", linetype="dashed")
p1<-p1+xlab("Fitted values")+ylab("Residuals")
p1<-p1+ggtitle("Rent: Residual vs Fitted Plot")+theme_bw()
p2<-ggplot(lm.rentsqm, aes(.fitted, .resid))+geom_point()
p2<-p2+stat_smooth(method="loess")+geom_hline(yintercept=0, col="red", linetype="dashed")
p2<-p2+xlab("Fitted values")+ylab("Residuals")
p2<-p2+ggtitle("rentsqm: Residual vs Fitted Plot")+theme_bw()
ggarrange(p1, p2, ncol=2)
```



Take home message: for the mean of the response may differ with out covariates - that is why we use regression. For the normal linear regression it is not the response that is supposed to have mean zero, but the error term - more about this in Module 2. And, is the variance of the residuals independent of the fitted values? Yes, more in Module 2.

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 to look for answers to your R questions (but also try the GLM teaching team)