Advanced dataanalysis and statistical modelling.

Introduction

Jan Kloppenborg Møller, Henrik Madsen
Poul Thyregod

DTU Compute
Technical University of Denmark
DK-2800 Kgs. Lyngby

January 2018

This lecture

- Introduction to the book
- Examples of types of data
- Motivating examples
- A first view on the models

Practical info.

- 4 hours every Monday (13.00-17.00)
- app. 2 hours of lecture and 2 hours of group exercise
- Grades will be based on 3 individual reports
- Handed in through campusnet (see handin dates there)
- You may work in groups but you should hand in seperate (and unique) repots
- You may use your own data for the projects, in this case you should discuss this with me.

Overview

The book

- The book provides an introduction to methods for statistical modeling using essentially all kind of data.
- The principles for modeling are based on likelihood techniques.
- Each chapter of the book contains examples and guidelines for solving the problems using the statistical software package R.
- The focus is on establishing models that explain the variation in data in such a way that the obtained models are well suited for predicting the outcome for given values of some explanatory variables.
- Focus on *formulating*, *estimating*, *validating and testing models* for predicting the *mean value* of the random variables.
- Consider the complete stochastic model for the data which includes an appropriate choice of the *density* describing the variation of the data.

The book

- Methods for modelling Gaussian distributed data, *regression analysis*, *analysis of variance* and the *analysis of covariance*, are established so that extension to similar methods applied in the case of, e.g. Poisson, Gamma and Binomial distributed data is easy using the likelihood approach in both cases.
- *General linear models* are relevant for *Gaussian distributed samples* whereas the *generalized linear models* facilitate a modeling of data originating from the so-called *exponential family of densities* including Poisson, Binomial, Exponential, Gaussian, and Gamma distributions.
- The presentation of the general and generalized linear models is provided using essentially the same methods related to the likelihood principles, but described in two separate chapters.
- The book also contains a first introduction to both mixed effects models (also called mixed models) and hierarchical models.

Notation

- All vectors are column vectors.
- Vectors and matrices are emphasized using a bold font.
- Lowercase letters are used for vectors and uppercase letters are used for matrices.
- Transposing is denoted with the upper index ^{T}.
- Random variables are always written using uppercase letters.
- Variables and random variables are assigned to letters from the last part of the alphabet (X, Y, Z, U, V, ...), while constants are assigned to letters from the first part of the alphabet (A, B, C, D, ...).
- From the context it should be possible to distinguish between a matrix and a random vector.

Overview

Types of data

1. *Continuous data* (e.g. $y_1 = 2.3$, $y_2 = -0.2$, $y_3 = 1.8$, $\ldots$, $y_n = 0.8$). Normal (Gaussian) distributed. Used, e.g. for air temperatures in degrees Celsius.
2. *Continuous positive data* (e.g. $y_1 = 0.0238$, $y_2 = 1.0322$, $y_3 = 0.0012$, $\ldots$, $y_n = 0.8993$). Log-normally distributed. Often used for concentrations.
3. *Count data*(e.g. $y_1 = 57$, $y_2 = 67$, $y_3 = 54$, $\ldots$, $y_n = 59$). Poisson distributed. Used, e.g. for number of accidents.
4. *Binary (or quantal) data* (e.g. $y_1 = 0$, $y_2 = 0$, $y_3 = 1$, $\ldots$, $y_n = 0$), or proportion of counts (e.g. $y_1 = 15/297$, $y_2 = 17/242$, $y_3 = 2/312$, $\ldots$, $y_n = 144/285$). Binomial distribution.
5. *Nominal data* (e.g. "Very unsatisfied", "Unsatisfied", "Neutral", "Satisfied", "Very satisfied"). Multinomial distribution.

Overview

The Challenger disaster

On January 28, 1986, Space Shuttle Challenger broke apart 73 seconds into its flight and the seven crew members died. The disaster was due to a disintegration of an O-ring seal in the right rocket booster. The forecast for January 28, 1986 indicated an unusually cold morning with air temperatures around 28 degrees F (-1 degrees C).

The planned launch on January 28, 1986 was launch number 25. During the previous 24 launches problems with the O-ring were observed in 6 cases. A model of the probability for O-ring failure as a function of the air temperature would clearly have shown that given the forecasted air temperature, problems with the O-rings were very likely to occur.

The Challenger disaster

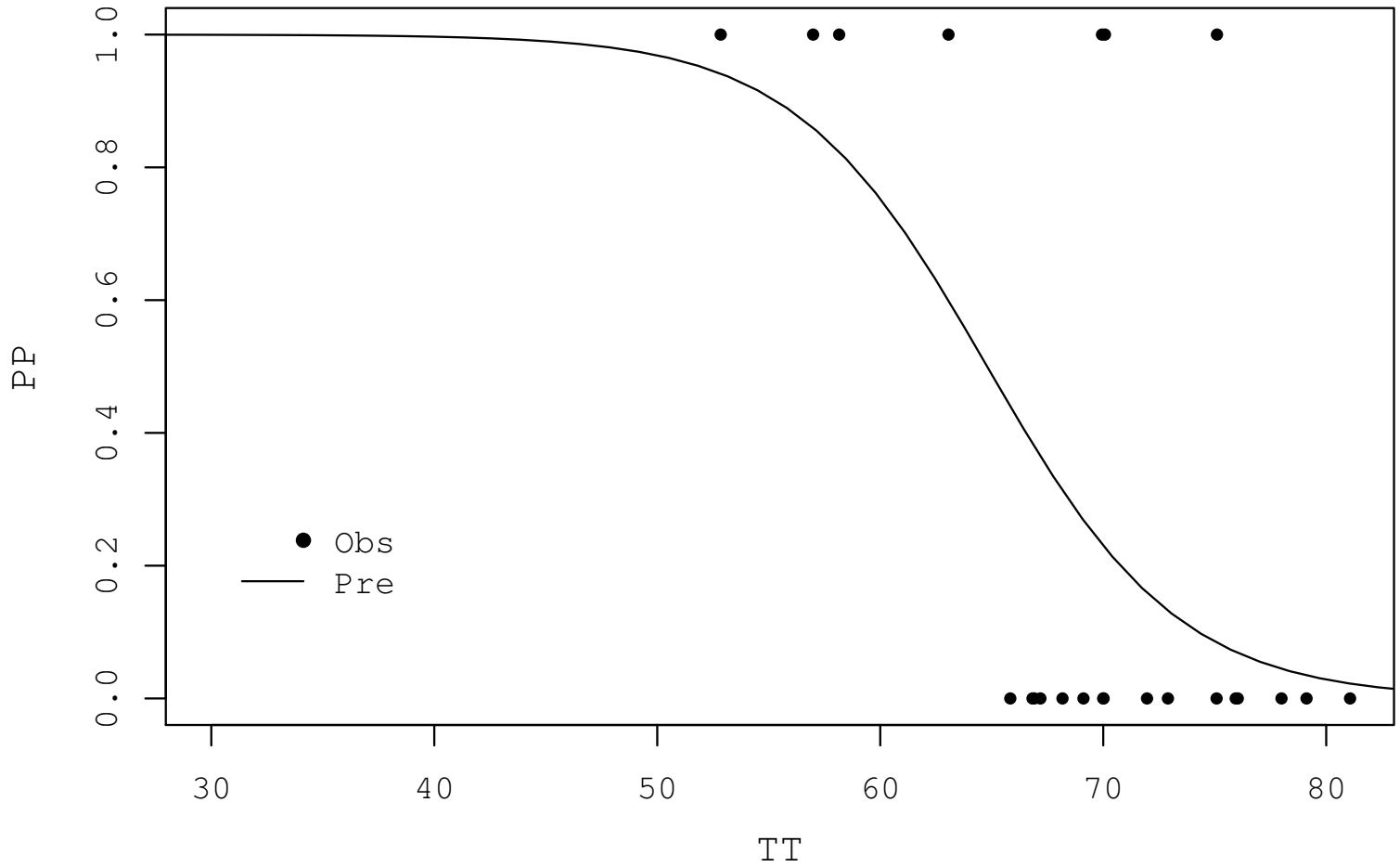


Figure: Observed failure of O-rings in 6 out of 24 launches along with predicted probability for O-ring failure.

QT prolongation for drugs

In the process of drug development it is required to perform a study of potential prolongation of a particular interval of the electrocardiogram (ECG), the QT interval. The QT interval is defined as the time required for completion of both ventricular depolarization and repolarization. The interval has gained clinical importance since a prolongation has been shown to induce potentially fatal ventricular arrhythmia such as Torsade de Pointes (TdP).

A number of drugs have been reported to prolong the QT interval, both cardiac and non-cardiac drugs. Recently, both previously approved as well as newly developed drugs have been withdrawn from the market or have had their labeling restricted because of indication of QT prolongation.

QT prolongation for drugs

Below are the results from a clinical trial where a QT prolonging drug was given to high risk patients. The patients were given the drug in six different doses and the number of incidents of Torsade de Points counted.

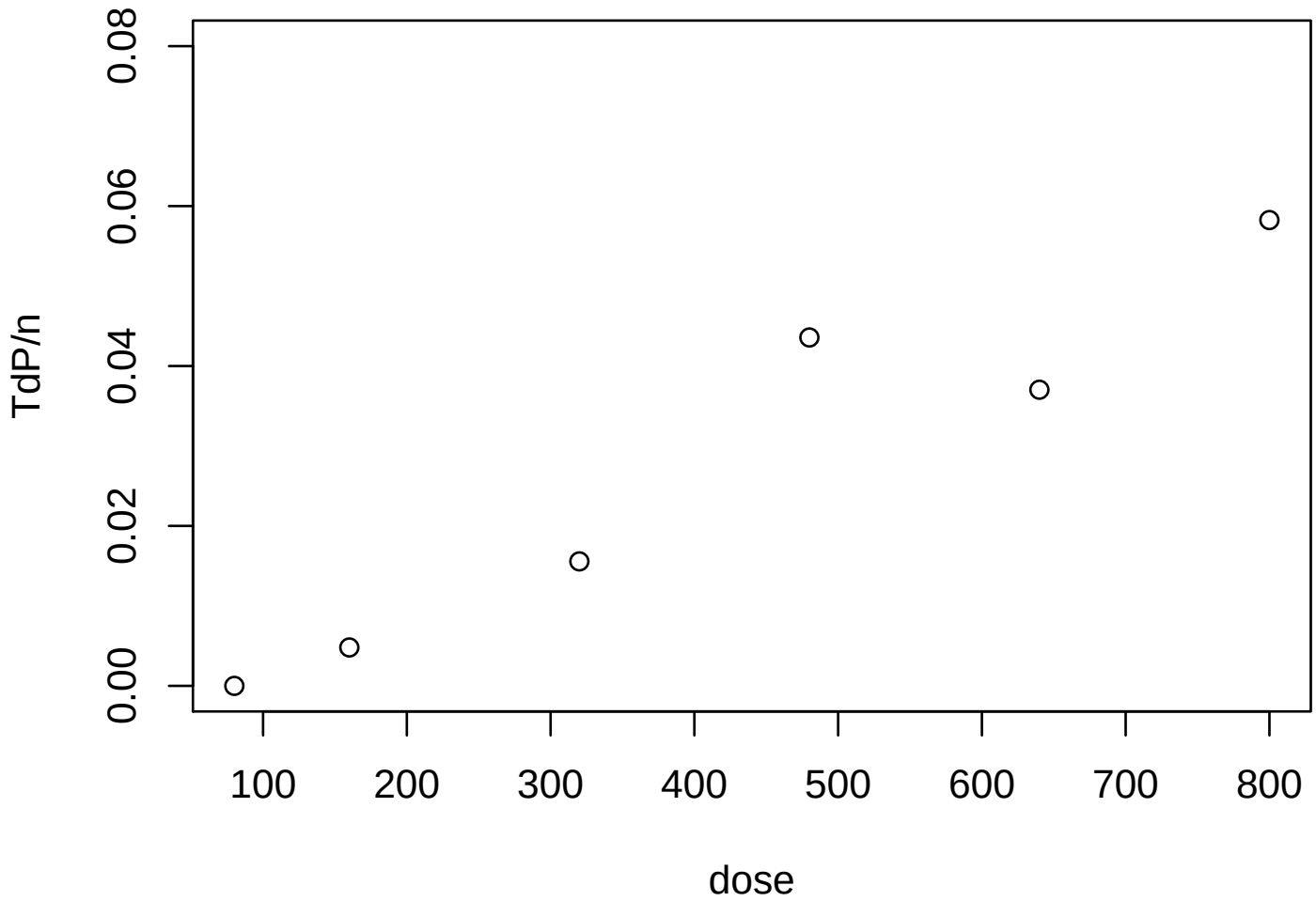
Index	Daily dose [mg]	Number of subjects	Number showing TdP	Fraction showing TdP
i	x_i	n_i	z_i	p_i
1	80	69	0	0
2	160	832	4	0.5
3	320	835	13	1.6
4	480	459	20	4.4
5	640	324	12	3.7
6	800	103	6	5.8

Table: Incidence of Torsade de Pointes by dose for high risk patients.

QT prolongation for drugs

- It is reasonable to consider the *fraction*, $Y_i = \frac{Z_i}{n_i}$, of incidences of Torsade de Points as the interesting variable.
- A natural distributional assumption is the binomial distribution, $Y_i \sim B(n_i, p_i)/n_i$, where n_i is the number of subjects given the actual dosage and p_i is the fraction showing Torsade de Pointes.

QT Prolongation



QT prolongation for drugs - bad model

- The fraction, p_i is higher for a higher daily dosage of the drug.
- A linear model of the form $Y_i = p_i + \epsilon_i$ where $p_i = \beta_0 + \beta_1 x_i$ does not reflect that p_i is between zero and one and the model for the fraction, Y_i (as "mean plus noise") is clearly not adequate, since the observations are between zero and one.
- It is clear that the distribution of ϵ_i and then the variance of observations must be dependent on p_i.
- Also, the problem with the homogeneity of the variance indicates that a traditional ("mean plus noise") model is not adequate here.

QT prolongation for drugs - correct model

Instead we will now formulate a model for transformed values of the observed fractions p_i.

Given that $Y_i \sim B(n_i, p_i)/n_i$ we have that

$$E[Y_i] = p_i$$

$$Var[Y_i] = \frac{p_i(1-p_i)}{n_i}$$

i.e. the variance is now a function of the mean value. Later on the so-called mean value function $V(E[Y_i])$ will be introduced which relates the variance to the mean value.

QT prolongation for drugs - correct model

We will consider a function, the so-called *link function* of the mean value $E[Y]$. In this case we will use the *logit*-transformation

$$g(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

and we will formulate a *linear model* for the transformed values.

QT prolongation for drugs - correct model

A plot of the observed logits, $g(p_i)$ as a function of the concentration indicates a linear relation of the form

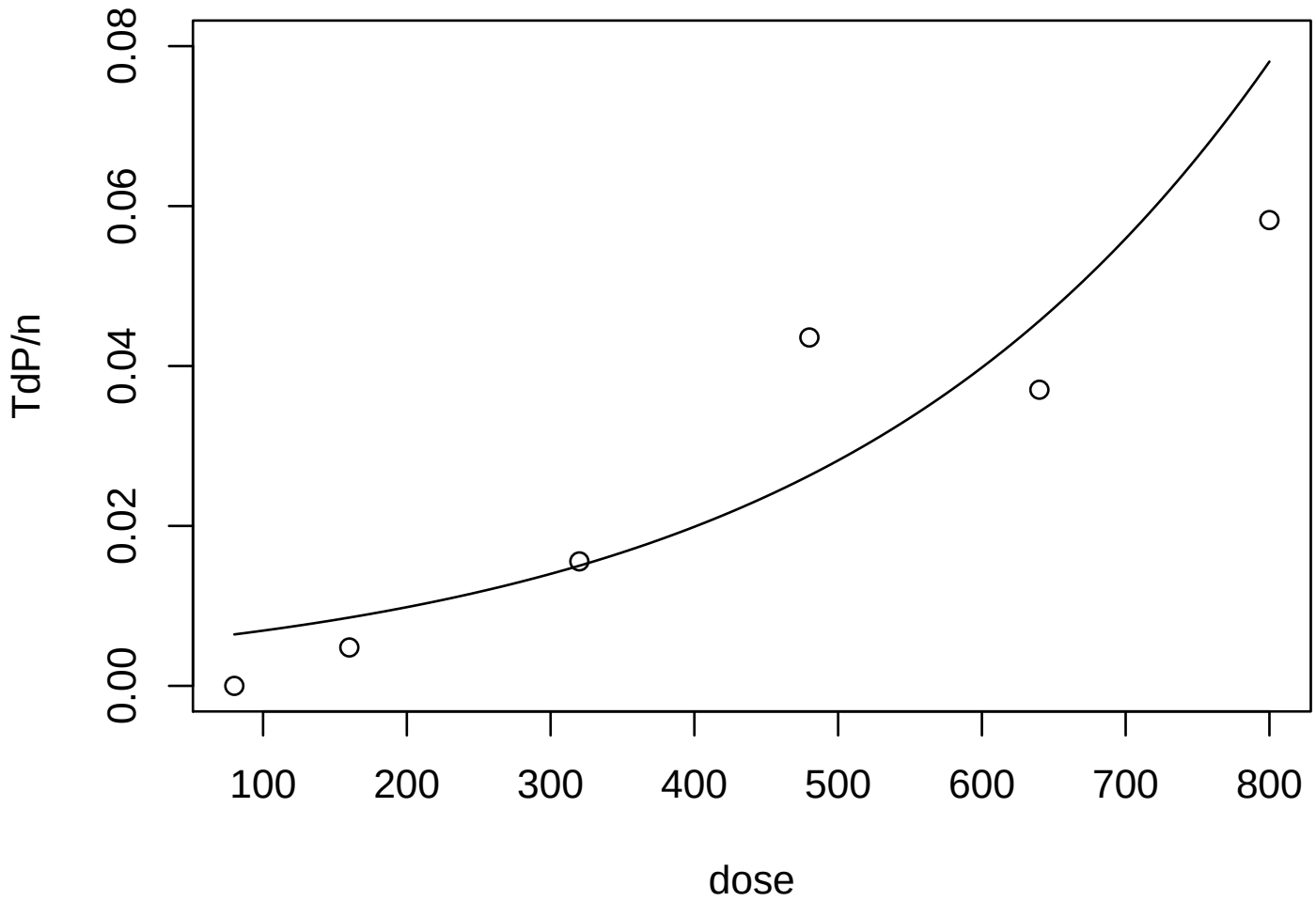
$$g(p_i) = \beta_0 + \beta_1 x_i$$

After having estimated the parameters, it is now possible to use the inverse transformation, which gives the predicted fraction $\widehat{p}$ of subjects showing Torsade de Pointes as a function of a daily dose, x using the *logistic function*:

$$\widehat{p} = \frac{\exp\left(\widehat{\beta}_0 + \widehat{\beta}_1 x\right)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 x)}.$$

This approach is called *logistic regression*.

QT Prolongation



Overview

A first view on the models

- We will focus on statistical methods to formulate models for predicting the expected value of the *outcome*, *dependent*, or *response variable*, Y_i as a function of the known *independent variables*, $x_{i1}, x_{i2}, \ldots, x_{ik}$.
- These k variables are also called *explanatory*, or *predictor variables* or *covariates*.
- This means that we shall focus on models for the expectation $E[Y_i]$.

A first view on the models

- Examples of types of response variables was shown on slide 9.
- Also the explanatory variables might be labeled as *continuous*, *discrete*, *categorical*, *binary*, *nominal*, or *ordinal*.
- To predict the response, a typical model often includes a combination of such types of variables.
- Since we are going to use a likelihood approach, a specification of the probability distribution of Y_i is a very important part when specifying the model.

General linear models

In *general linear models*, the expected value of the response variable Y is linked linearly to the explanatory variables by an equation of the form

$$E[Y_i] = \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \ .$$

It will be shown that for Gaussian data it is reasonable to build a model directly for the expectation. This relates to the fact that for Gaussian distributed random variables, all conditional expectations are linear.

Generalized linear models

It is often more reasonable to build a linear model for a transformation of the expected value of the response. This approach is more formally described in connection with the *generalized linear models* where a link between the expected value of response and the explanatory variables is of the form

$$g(E[Y_i]) = \beta_1 x_{i1} + \ldots + \beta_k x_{ik} \ .$$

The function $g(.)$ is called the *link function* and the right hand side of the equation is called the *linear component* of the model.

Generalized linear models

A full specification of the model contains a specification of

1. The *probability density* of Y. In the general linear model this will be the Gaussian density, i.e. $Y \sim N(\mu, \sigma^2)$, whereas in the generalized linear model the probability density will belong to the *exponential family of densities*, which includes the Gaussian, Poisson, Binomial, Gamma, and other distributions.
2. The smooth monotonic *link function* $g(.)$. Here we have some freedom, but the so-called *canonical link* function is directly linked to the used density. No link function is needed for Gaussian data – or the link is the identity.
3. The *linear component*.

Hierarchical models

- In Chapters 5 and 6 of the book the important concept of *hierarchical models* is introduced.
- The Gaussian case is introduced in Chapter 5, and this includes the so-called linear mixed effects models.
- This Gaussian and linear case is a natural extension of the general linear models.
- An extension of the generalized linear models are found in Chapter 6 which briefly introduces the generalized hierarchical models.

Hierarchical models - Gaussian case

Consider for instance the test of ready made concrete. The concrete are delivered by large trucks. From a number of randomly picked trucks a small sample is taken, and these samples are analyzed with respect to the strength of concrete. A reasonable model for the variation of the strength is

$$Y_{ij} = \mu + U_i + \epsilon_{ij}$$

where μ is the overall strength of the concrete and U_i is the deviation of the average for the strength of concrete delivered by the i'th truck, and $\epsilon_{ij} \sim N(0, \sigma^2)$ the deviation between concrete samples from the same truck.
Here we are typically not interested in the individual values of U_i but rather in the variation of U_i, and we will assume that $U_i \sim N(0, \sigma_u^2)$.

Hierarchical models - Gaussian case

The model on slide 29 is a *one-way random effects model*. The parameters are now μ, σ_u^2 and σ^2.
Putting $\mu_i = \mu + U_i$ we may formulate the model as a *hierarchical model*, where we shall assume that

$$Y_{ij}|\mu_i \sim N(\mu_i, \sigma^2) \ ,$$

and in contrast to the *fixed effects model*, the level μ_i is modeled as a realization of a random variable,

$$\mu_i \sim N(\mu, \sigma_u^2),$$

where the μ_i's are assumed to be mutually independent, and Y_{ij} are *conditionally independent*, i.e. Y_{ij} are mutually independent in the conditional distribution of Y_{ij} for given μ_i.

Hierarchical models - Gaussian case

Let us again consider a model for all n observations and let us further extend the discussion to the vector case of the random effects. The discussion above can now be generalized to the *linear mixed effects model* where

$$E[\boldsymbol{Y}|\boldsymbol{U}] = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{U}$$

with $\boldsymbol{X}$ and $\boldsymbol{Z}$ denoting known matrices. Note how the mixed effect linear model in is a linear combination of *fixed effects*, $\boldsymbol{X}\boldsymbol{\beta}$ and *random effects*, $\boldsymbol{Z}\boldsymbol{U}$. These types of models will be described in Chapter 5.

Hierarchical models - non-Gaussian case

The non-Gaussian case of the hierarchical models, where

$$g(E[\boldsymbol{Y}|\boldsymbol{U}]) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{U}$$

and where $g(.)$ is an appropriate link function will be treated in Chapter 6.

Overview

The Maximum Likelihood estimate

We will assume that the statistical model for $\boldsymbol{y}$ is given by parametric family of joint densities:

$$\{f_{\mathbf{Y}}(y_1, y_2, \ldots, y_n; \boldsymbol{\theta})\}_{\boldsymbol{\theta} \in \Theta^k}$$

Remember that when the n random variables are independent, the joint probability density equals the product of the corresponding marginal densities or:

$$f(y_1, y_2, ...y_n) = f_1(y_1) \cdot f_2(y_2) \cdot \ldots \cdot f_n(y_n)$$

Definition (Maximum Likelihood Estimate (MLE))

Given the observation $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ the *Maximum Likelihood Estimate (MLE)* is a function $\widehat{\boldsymbol{\theta}}(\boldsymbol{y})$ such that

$$L(\widehat{\boldsymbol{\theta}}; \boldsymbol{y}) = \sup_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}; \boldsymbol{y}) = \sup_{\boldsymbol{\theta} \in \Theta} f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})$$

Likelihood in this course

- Likelihood theory form the basis of most of what we do in this course, and an overview is given in lecture 2 next week.
- If you are not familiar with likelihood estimation, you should study Chapter 2 of the text book before next lecture

Overview

1. Introduction to the book
2. Examples of types of data
3. Motivating examples
4. A first view on the models
5. Likelihood estimation