Introduction to Statistics

Dietrich von Rosen[1]

September 2, 2011

[1]Department of Engineering and Technology, Swedish University of Agricultural Sciences (dietrich.von.rosen@et.slu.se).

1 Introduction

1.1 What is statistics

Statistical sciences is about planning experiments, setting up models to analyze experiments and to study properties of these models or study properties of some specific building blocks within these models, e.g. parameters and independence assumptions. Statistical application is about connecting statistical models to data. The general statistical paradigm constitutes of the following steps

1. Set up a model,
2. Evaluate the model via simulations or comparisons with data,
3. If necessary refine the model and restart from step 2,
4. Accept and interpret the model.

It is obvious that a number of decisions have to be made within this paradigm which unfortunately all are rather subjective. This should be taken into account when relying on statistics. Moreover, in order that statistics should become useful the model should be relevant for the problem under consideration which is often relative to the information available from data, and the final model should be interpretable. Statistics is instrumental since one usually can not draw firm conclusions without expertise about the data which is used to evaluate the model. On the other hand "data experts" when applying statistics need a solid knowledge in statistics to perform efficient analysis.

Basic ingredients in statistics is the concept of probability and the assumption about underlying distributions. The distribution is a probability measure on the space of observations but what is a probability and what does a probability represent? Statistics uses the concept of probability as a measure of uncertainty. The probability measure is well defined through its characterization via Kolmogorov's axioms, although there are discussions concerning the definition of conditional probabilities. However, Kolmogorov's axioms tell us what a probability measure should fulfill but not what it is. It is even not obvious that in real life (Nature) something like a probabilistic mechanism exists but for statisticians this does not matter. The probability measure is part of a model and any model only describes reality approximatively.

1.2 What is a statistical model

A statistical model is usually a class of distributions which is specified via relationships on parameters (unknown quantities). The idea is to choose an appropriate model class according to the problem which will be studied. Sometimes we know exactly what distribution should be used but more often we have parameters which generate a model class. For example, the class of multivariate normal distributions of fixed size but with unknown mean and dispersion. Instead of distributions it may be convenient, in particular for interpretations, to work with random variables which are representatives of the random phenomenon under study. The problem with statistics is how to connect data to continuous random variables. In general it has become fruitful to look upon data as realizations

of random variables. However, since our data points have probability mass 0 we can not couple, in a mathematical way, continuous random variables to data.

There exist several well known schools in statistics of how to connect data in a more or less rigorous way to statistical models. For example, "Distribution free" methods, Likelihood based methods, and Bayesian methods. The approaches have been ordered with respect to how much they relay on distributional properties to various depths. Observe that Distribution free method does not mean that there is no assumption about any model. In a statistical model there are always some assumptions about randomness, for example concerning independence between random variables. Maybe the most well known Distribution free method is the least squares approach.

Likelihood methods utilizes distributions where the classes of distributions are generated by unknown parameters and the idea is to estimate these parameters. A consequence of this procedure is that we obtain which distribution we should consider as the true distribution as well as we get information about the parameters which if the model is appropriately specified are interpretable. Concerning the normal distribution usually the mean and variance act as parameters, although from an exponential family point of view a more natural equivalent parametrization can be set up.

In Bayesian methods the basic idea is that everything unknown is modelled with the help of distributions, among others parameters. One is avoiding some of the problems with the likelihood approach such, as connecting continuous data to a model, but instead one generates others, for example it is difficult to specify distributions for all unknown elements. Moreover, in the Bayesian approach the concept of conditional independence is crucial in contrary to the Likelihood approach where independence is used. Which to prefer is a matter of taste.

Example 1.1. *(Several statistical approaches for evaluating a univariate liner model.)* Let

$$\boldsymbol{x}' = \boldsymbol{\beta}'\boldsymbol{C} + \boldsymbol{e}',$$

where $\boldsymbol{x} : n \times 1$, a random vector corresponding to the observations, $\boldsymbol{C} : k \times n$, the design matrix, $\boldsymbol{\beta} : k \times 1$ is an unknown parameter vector which is to be estimated, and $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ which is considered to be the error term in the model. The least squares approach works as follows: Let $\boldsymbol{x}_0$ be the observations and we will minimize with respect to $\boldsymbol{\beta}$

$$(\boldsymbol{x}_0' - \boldsymbol{\beta}'\boldsymbol{C})(\boldsymbol{x}_0' - \boldsymbol{\beta}'\boldsymbol{C})'$$

which gives

$$\widehat{\boldsymbol{\beta}}_0'\boldsymbol{C} = \boldsymbol{x}_0'\boldsymbol{P}_{C'}$$

since

$$(\boldsymbol{x}_0' - \boldsymbol{\beta}'\boldsymbol{C})(\boldsymbol{x}_0' - \boldsymbol{\beta}'\boldsymbol{C})' = \boldsymbol{x}_0'(\boldsymbol{I} - \boldsymbol{P}_{C'})\boldsymbol{x}_0 + (\boldsymbol{P}_{C'} - \boldsymbol{\beta}'\boldsymbol{C})()',$$

where $\boldsymbol{\beta}_0$ stands for the estimate of $\boldsymbol{\beta}$, i.e. an explicit numerical value of $\boldsymbol{\beta}$ and the projector $\boldsymbol{P}_{C'} = \boldsymbol{C}'(\boldsymbol{C}\boldsymbol{C}')^{-}\boldsymbol{C}$. In order to study properties $\boldsymbol{x}_0$ is replaced by $\boldsymbol{x}$ and doing so we get the estimator

$$\widehat{\boldsymbol{\beta}}'\boldsymbol{C} = \boldsymbol{x}'\boldsymbol{P}_{C'}.$$

Due to the linearity of the estimator

$$\widehat{\boldsymbol{\beta}}'\boldsymbol{C} \sim N_n(\boldsymbol{\beta}\boldsymbol{C}, \sigma^2\boldsymbol{P}_{C'}),$$

i.e. $\widehat{\boldsymbol{\beta}}\boldsymbol{C}$ is unbiased and normally distributed with variance $\sigma^2\boldsymbol{P}_{C'}$. Moreover, the variance parameter may be estimated as $n\widehat{\sigma}^2 = \boldsymbol{x}'(\boldsymbol{I} - \boldsymbol{P}_{C'})\boldsymbol{x}$. The model may among others be evaluated via residuals, i.e. $\boldsymbol{x}_0'(\boldsymbol{I}-\boldsymbol{P}_{C'})$ and $\boldsymbol{x}'(\boldsymbol{I}-\boldsymbol{P}_{C'})$. For example one should evaluate the model with respect to influential observations and outliers as well as the fit of the model to data. Moreover, specific properties such as smallest variance properties of the estimator may be shown or best quadratic properties of the variance estimator. □

An alternative estimation procedure is based on finding estimators which minimize the overall variance

$$E[(\boldsymbol{x}' - \widehat{\boldsymbol{\beta}}'\boldsymbol{C})(\boldsymbol{x}' - \widehat{\boldsymbol{\beta}}'\boldsymbol{C})']$$

which can be manipulated in the following way

$$\begin{aligned} &E[(\boldsymbol{x}' - \widehat{\boldsymbol{\beta}}'\boldsymbol{C})(\boldsymbol{x}' - \widehat{\boldsymbol{\beta}}'\boldsymbol{C})'] \\ &\quad = \quad E[\boldsymbol{x}'(\boldsymbol{I} - \boldsymbol{P}_{C'})\boldsymbol{x}] + E[(\boldsymbol{x}'\boldsymbol{P}_{C'} - \widehat{\boldsymbol{\beta}}'\boldsymbol{C})(\boldsymbol{x}'\boldsymbol{P}_{C'} - \widehat{\boldsymbol{\beta}}'\boldsymbol{C})']. \end{aligned}$$

Thus, it follows that the estimator equals

$$\widehat{\boldsymbol{\beta}}'\boldsymbol{C} = \boldsymbol{x}'\boldsymbol{P}_{C'}.$$

In order to verify the model via comparisons to data the estimate

$$\widehat{\boldsymbol{\beta}}_0'\boldsymbol{C} = \boldsymbol{x}_0'\boldsymbol{P}_{C'}$$

is calculated. These expressions are all the same as for the least squares approach, although the methods conceptually differ a lot, i.e. for the least squares method we start with data, find an estimate, and then construct an estimator by replacing the data, $\boldsymbol{x}_0$, with $\boldsymbol{x}$. For the minimization of the variance we started with $\boldsymbol{x}$, found an estimator, and then constructed an estimate by replacing $\boldsymbol{x}$ by $\boldsymbol{x}_0$.

Now we turn to the likelihood approach. Here one starts with the likelihood which is the density of $\boldsymbol{x}$ evaluated at $\boldsymbol{x}_0$, i.e.

$$L(\boldsymbol{\beta}, \sigma^2) = (2\pi)^{-n/2}(\sigma^2)^{n/2}\exp\{-\tfrac{\sigma^2}{2}(\boldsymbol{x}_0' - \boldsymbol{\beta}'\boldsymbol{C})(\boldsymbol{x}_0' - \boldsymbol{\beta}'\boldsymbol{C})'\}.$$

This function is maximized with respect to σ^2 and $\boldsymbol{\beta}$ which gives

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_0'\boldsymbol{C} &= \boldsymbol{x}_0'\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{C}')^{-}\boldsymbol{C}, \\ n\widehat{\sigma}_0^2 &= \boldsymbol{x}_0'(\boldsymbol{I} - \boldsymbol{C}'(\boldsymbol{C}\boldsymbol{C}')^{-}\boldsymbol{C})\boldsymbol{x}_0. \end{aligned}$$

In the next step of the likelihood approach $\boldsymbol{\beta}$ is constructed by replacing $\boldsymbol{x}_0$ by $\boldsymbol{x}$. Hence we have that the likelihood approach will lead to the same conclusion as the least squares approach and the minimum variance methods.

In the Bayesian approach the idea is to evaluate the posterior density (probability) function for, say $\boldsymbol{\beta}$, i.e. the density function for $\boldsymbol{\beta}$ given the data. If supposing that σ^2 is known we only have to consider $\boldsymbol{\beta}$ and thus one studies

$$f(\boldsymbol{\beta}|\boldsymbol{x}) = \frac{f(\boldsymbol{\beta}, \boldsymbol{x})}{f(\boldsymbol{x})} = \frac{f(\boldsymbol{x}|\boldsymbol{\beta})f(\boldsymbol{\beta})}{f(\boldsymbol{x})}.$$

Note that $f(\boldsymbol{x}|\boldsymbol{\beta})$ is the likelihood function and it remains to specify $f(\boldsymbol{\beta})$, the a priori density for $\boldsymbol{\beta}$. Let $\boldsymbol{\beta} \sim N_k(\boldsymbol{\beta}_0, \boldsymbol{I})$ and then (suppose $\boldsymbol{C}$ is of full rank).

$$\begin{aligned} f(\boldsymbol{\beta}|\boldsymbol{x} = \boldsymbol{x}_0) \quad \propto \quad & \exp\{-\tfrac{1}{2\sigma^2}(\boldsymbol{x}_0'\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{C}')^{-1} - \boldsymbol{\beta}')\boldsymbol{C}\boldsymbol{C}'(\boldsymbol{x}_0'\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{C}')^{-1} - \boldsymbol{\beta}')'\} \\ & \times \exp\{-\tfrac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'\}. \end{aligned}$$

With the help of $f(\boldsymbol{\beta}|\boldsymbol{x} = \boldsymbol{x}_0)$ the vector $\boldsymbol{\beta}$ should be evaluated. There are many approaches available but one simple is to consider the posterior mean $E[\boldsymbol{\beta}|\boldsymbol{x} = \boldsymbol{x}_0]$. Performing some calculation gives

$$E[\boldsymbol{\beta}|\boldsymbol{x} = \boldsymbol{x}_0] = \boldsymbol{x}_0'\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{C}')^{-1} + \boldsymbol{\beta}_0.$$

□

We would also like to emphasize that models should be understandable, i.e. parameters and their estimators should be understandable, and computations should be quick. For the last 20 years entering the world of PCs and strong computer facilities has led to absurd use of algorithms and one can see programs running days and nights. The beauty of Statistics as well as its relation to Mathematics has been partly lost. This is serious because mathematics helps us to look through the models and helps us to understand the analysis. Without mathematics it is easy to become trapped in too many ad-hoc procedures. Intuition and ad-hoc procedures should be basic ingredients in statistical model building but they should also be possible to verify. This is the best way to produce something which later may be improved. To use too many simulation studies will end up into something which only with difficulties can be transmitted to the next generation of statisticians.

1.3 The General Univariate Linear Model with Known Dispersion

In this section the classical Gauss-Markov set up will be considered but we assume the dispersion matrix to be known. If the dispersion matrix is p.d. the model is just a minor extension of the model in Example 1.1. However, if the dispersion matrix is p.s.d. other aspects related to the model will be introduced. In general, in the Gauss-Markov model the dispersion is proportional to an unknown constant but this is immaterial for our presentation. The reason for investigating the model in some detail is that there has to be a close connection between the estimators based on models with known and unknown dispersion.

Later on all our models if assuming a known dispersion matrix can be reformulated as a Gauss-Markov model. With the additional information that the random variables are normally distributed, one can see this from the likelihood equations. Moreover, it follows from these equations that the maximum likelihood estimators of the mean parameters under the assumption of unknown dispersion should approach the estimators with known dispersion. For example the likelihood equation for the model $\boldsymbol{X} \sim N_{p,n}(\boldsymbol{ABC}, \boldsymbol{\Sigma}, \boldsymbol{I})$ which appears when differentiating with respect to $\boldsymbol{B}$ equals

$$\boldsymbol{A}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{ABC})\boldsymbol{C}' = \boldsymbol{0}$$

and for a large sample any MLE of $\boldsymbol{B}$ has to asymptotically satisfy this equation because we know that the MLE of $\boldsymbol{\Sigma}$ is consistent.

Now let

$$\boldsymbol{x}' = \boldsymbol{\beta}'\boldsymbol{C} + \boldsymbol{e}', \qquad \boldsymbol{e} \sim N_n(\boldsymbol{0}, \boldsymbol{V}), \tag{1.1}$$

where $\boldsymbol{V} : n \times n$ is p.d. and known, $\boldsymbol{x} : n \times 1$, $\boldsymbol{C} : k \times n$ and $\boldsymbol{\beta} : k \times 1$ is to be estimated. Let $\boldsymbol{x}_0$ as previously denote the observations of $\boldsymbol{x}$. Then, the likelihood is maximized as follows (use that $\boldsymbol{V}^{-1} = \boldsymbol{V}^{-1}\boldsymbol{P}_{C',V^{-1}} + \boldsymbol{P}_{C^{o\prime},V}\boldsymbol{V}^{-1}$)

$$\begin{aligned} L(\boldsymbol{\beta}) &\propto |\boldsymbol{V}|^{-1/2}\exp\{-1/2(\boldsymbol{x}_0' - \boldsymbol{\beta}'\boldsymbol{C})\boldsymbol{V}^{-1}(\boldsymbol{x}_0' - \boldsymbol{\beta}'\boldsymbol{C})'\} \\ &= |\boldsymbol{V}|^{-1/2}\exp\{-1/2(\boldsymbol{x}_0'\boldsymbol{P}'_{\mathrm{C}',\mathrm{V}^{-1}} - \beta'\boldsymbol{C})\boldsymbol{V}^{-1}()'\} \\ &\quad \times\exp\{-1/2(\boldsymbol{x}_0'\boldsymbol{P}_{\mathrm{C}^{o\prime},\mathrm{V}}\boldsymbol{V}^{-1}\boldsymbol{x}_0)\} \\ &\leq |\boldsymbol{V}|^{-1/2}\exp\{-1/2(\boldsymbol{x}_0'\boldsymbol{P}_{\mathrm{C}^{o\prime},\mathrm{V}}\boldsymbol{V}^{-1}\boldsymbol{x}_0)\}, \end{aligned}$$

which is independent of any parameter, i.e. $\boldsymbol{\beta}$, and the upper bound is obtained iff

$$\widehat{\boldsymbol{\beta}}_0'\boldsymbol{C} = \boldsymbol{x}_0'\boldsymbol{P}'_{C',V^{-1}},$$

where $\widehat{\boldsymbol{\beta}}_0$ is the estimate of $\boldsymbol{\beta}$. Thus, in order to estimate $\boldsymbol{\beta}$ a linear equation system has to be solved. The solution can be written

$$\widehat{\boldsymbol{\beta}}_0' = \boldsymbol{x}_0'\boldsymbol{V}^{-1}\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{V}^{-1}\boldsymbol{C}')^- + \boldsymbol{z}'(\boldsymbol{C})^{o'},$$

where $\boldsymbol{z}'$ stands for an arbitrary vector.

Suppose now that in model (1.1) we have restrictions on the mean vector given by

$$\boldsymbol{\beta}'\boldsymbol{G} = \boldsymbol{0}.$$

Then

$$\boldsymbol{\beta}' = \boldsymbol{\theta}'\boldsymbol{G}^{o'}$$

where $\boldsymbol{\theta}$ is a new unrestricted parameter. After inserting this relation in (1.1) the following model appears:

$$\boldsymbol{x}' = \boldsymbol{\theta}'\boldsymbol{G}^{o'}\boldsymbol{C} + \boldsymbol{e}', \qquad \boldsymbol{e} \sim N_n(\boldsymbol{0}, \boldsymbol{V}).$$

Thus, the above presented calculations yield

$$\widehat{\boldsymbol{\beta}}_0'\boldsymbol{C} = \boldsymbol{x}_0'\boldsymbol{P}'_{C'\boldsymbol{G}^o,V^{-1}}$$

and from hear a general expression for $\widehat{\boldsymbol{\beta}}_0$ ($\widehat{\boldsymbol{\beta}}$) is obtained.

If $\boldsymbol{V}$ is p.s.d the likelihood does not exist and the model consists of a continuous and a discrete part. Because $\boldsymbol{V}$ is p.s.d. there exists a semi orthogonal matrix $\boldsymbol{H} : n \times r$, where $r = \text{rank}\boldsymbol{V}$ and $\boldsymbol{V} = \boldsymbol{H}\boldsymbol{H}'$, such that

$$\boldsymbol{H}^{o'}\boldsymbol{V} = \boldsymbol{0}.$$

Observe that we do not loose any "information" if a one-one transformation of $\boldsymbol{x}$ takes place. Then the estimation of $\boldsymbol{\beta}$ in (1.1) can equivalently be carried out via $\boldsymbol{x}'(\boldsymbol{H}, \boldsymbol{H}^o)$. Hence, with probability 1

$$\boldsymbol{x}_0'\boldsymbol{H}^o = \boldsymbol{\beta}'\boldsymbol{C}\boldsymbol{H}^o \tag{1.2}$$

and therefore we assume (consistency assumption) $\boldsymbol{H}^{o'}\boldsymbol{x}_0 \in \mathcal{C}(\boldsymbol{H}^{o'}\boldsymbol{C}')$ which is equivalent to $\boldsymbol{x}_0 \in \mathcal{C}(\boldsymbol{C}' : \boldsymbol{V})$. Thus, the data put restrictions on $\boldsymbol{\beta}$ which is a new feature in comparison with the case when $\boldsymbol{V}$ is of full rank. If this is meaningful depends on the problem under consideration. Moreover,

$$\boldsymbol{x}'\boldsymbol{H} = \boldsymbol{\beta}'\boldsymbol{C}\boldsymbol{H} + \widetilde{\boldsymbol{e}}, \qquad \widetilde{\boldsymbol{e}} \sim N_r(\boldsymbol{0}, \boldsymbol{H}'\boldsymbol{V}\boldsymbol{H}). \tag{1.3}$$

Equation (1.2) is linear in $\boldsymbol{\beta}$ and because of consistency

$$\boldsymbol{\beta}' = \boldsymbol{x}_0'\boldsymbol{H}^o(\boldsymbol{C}\boldsymbol{H}^o)^- + \boldsymbol{\theta}'(\boldsymbol{C}\boldsymbol{H}^o)^{o'}$$

where one may view $\boldsymbol{\theta}$ as a new set of unrestricted parameters. Inserting the solution into (1.3) yields

$$\boldsymbol{x}'\boldsymbol{H} = \boldsymbol{x}_0'\boldsymbol{H}^o(\boldsymbol{C}\boldsymbol{H}^o)^-\boldsymbol{C}\boldsymbol{H} + \boldsymbol{\theta}'(\boldsymbol{C}\boldsymbol{H}^o)^{o'}\boldsymbol{C}\boldsymbol{H} + \widetilde{\boldsymbol{e}}.$$

From earlier calculations we know that

$$\begin{aligned}\boldsymbol{\theta}_0' = {} & \boldsymbol{x}_0'(\boldsymbol{I} - \boldsymbol{H}^o(\boldsymbol{C}\boldsymbol{H}^o)^-\boldsymbol{C})\boldsymbol{H} \\ & \times(\boldsymbol{H}'\boldsymbol{V}\boldsymbol{H})^{-1}\boldsymbol{H}'\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{H}^o)^o((\boldsymbol{C}\boldsymbol{H}^o)^{o'}\boldsymbol{C}\boldsymbol{H}(\boldsymbol{H}'\boldsymbol{V}\boldsymbol{H})^{-1}\boldsymbol{H}'\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{H}^o)^o)^- \\ & +\boldsymbol{z}'((\boldsymbol{C}\boldsymbol{H}^o)^{o'}\boldsymbol{C}\boldsymbol{H})^{o'},\end{aligned} \tag{1.4}$$

where $\boldsymbol{z}$ is an arbitrary vector and then

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_0' = {} & \boldsymbol{x}_0'\boldsymbol{H}^o(\boldsymbol{C}\boldsymbol{H}^o)^- + \boldsymbol{x}_0'(\boldsymbol{I} - \boldsymbol{H}^o(\boldsymbol{C}\boldsymbol{H}^o)^-\boldsymbol{C})\boldsymbol{H} \\ & \times(\boldsymbol{H}'\boldsymbol{V}\boldsymbol{H})^{-1}\boldsymbol{H}'\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{H}^o)^o((\boldsymbol{C}\boldsymbol{H}^o)^{o'}\boldsymbol{C}\boldsymbol{H}(\boldsymbol{H}'\boldsymbol{V}\boldsymbol{H})^{-1}\boldsymbol{H}'\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{H}^o)^o)^-(\boldsymbol{C}\boldsymbol{H}^o)^{o'} \\ & +\boldsymbol{z}'((\boldsymbol{C}\boldsymbol{H}^o)^{o'}\boldsymbol{C}\boldsymbol{H})^{o'}(\boldsymbol{C}\boldsymbol{H}^o)^{o'}.\end{aligned}$$

If studying statistical properties of this estimate we should consider (remember (1.2))

$$\begin{aligned}\widehat{\boldsymbol{\beta}}' &= \boldsymbol{\beta}'\boldsymbol{C}\boldsymbol{H}^o(\boldsymbol{C}\boldsymbol{H}^o)^- \\ &\quad -\boldsymbol{\beta}'\boldsymbol{C}\boldsymbol{H}^o(\boldsymbol{C}\boldsymbol{H}^o)^-\boldsymbol{C}\boldsymbol{H}\boldsymbol{H}'\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{H}^o)^o((\boldsymbol{C}\boldsymbol{H}^o)^{o'}\boldsymbol{C}\boldsymbol{H}\boldsymbol{H}'\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{H}^o)^o)^-(\boldsymbol{C}\boldsymbol{H}^o)^{o'} \\ &\quad +\boldsymbol{x}'\boldsymbol{H}\boldsymbol{H}'\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{H}^o)^o((\boldsymbol{C}\boldsymbol{H}^o)^{o'}\boldsymbol{C}\boldsymbol{H}\boldsymbol{H}'\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{H}^o)^o)^-(\boldsymbol{C}\boldsymbol{H}^o)^{o'} \\ &\quad +\boldsymbol{z}'((\boldsymbol{C}\boldsymbol{H}^o)^{o'}\boldsymbol{C}\boldsymbol{H})^{o'}(\boldsymbol{C}\boldsymbol{H}^o)^{o'}\end{aligned}$$

and then assume some conditions so that the term including $\boldsymbol{z}$ will disappear. Moreover, in practise the condition $\boldsymbol{x}_0 \in \mathcal{C}(\boldsymbol{C}' : \boldsymbol{V})$ may not be satisfied and then a pretreatment of data has to take place, e.g. a projection of data on the space $\mathcal{C}(\boldsymbol{C}' : \boldsymbol{V})$.

Example 1.2. *(Singular Gauss-Markov model).* In an experiment where eating behavior of n diary cows was studied through administration of food one could keep the total amount of food fixed (say t) over 24h. During the 24h it was recorded how much each of the n cows was eating. Due to breeding and local environment the cows are correlated with a covariance matrix $\sigma^2\boldsymbol{V}$, where σ^2 is an unknown scale parameter. Since the cows are part of many feeding experiments the correlation between cows may be supposed to be known. The main idea is to relate the recorded values to various variables such as lactation, amount of produced milk and various variables measuring the quality of the milk. If the measurements are denoted x_{0i}, $i = 1, 2, \ldots, n$ and the other explanatory variables $c_1, c_2, \ldots c_k$ the following linear model may be set up:

$$x_i = \mu + \sum_{j=1}^{k} \beta_j c_{ji} + \epsilon_i$$

which in matrix notation equals

$$\boldsymbol{x}' = \boldsymbol{\beta}'\boldsymbol{C} + \boldsymbol{e}',$$

where $\boldsymbol{C} = (\mathbf{1}_n, \boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_k)'$, $\boldsymbol{c}_j = (c_{ji})$ and $\boldsymbol{e} \sim N_n(\mathbf{0}, \sigma^2\boldsymbol{V})$ where σ^2 is an unknown parameter. As an estimator of σ^2 we may use

$$(n - k - 1)\widehat{\sigma}^2 = (\boldsymbol{x}' - \widehat{\boldsymbol{\beta}}'\boldsymbol{C})()'.$$

Thus, if we are able to estimate $\boldsymbol{\beta}$ all parameters can be estimated.

The technical treatment of the model goes as follows. Observe that by making a one-one transformation of $\boldsymbol{x}$ there is no information loss. Thus, $\boldsymbol{x}$ will be premultiplied by $\mathbf{1}'$ and $\mathbf{1}^{o'}$. Note that $\boldsymbol{x}'\mathbf{1} = t$ implies that $\boldsymbol{V}\mathbf{1} = \mathbf{0}$ and

$$\boldsymbol{\beta}'\boldsymbol{C}\mathbf{1} = \boldsymbol{x}'\mathbf{1} = t.$$

This means that according to the model we have an equation with no variation and thus the equation can be treated as a deterministic equation which puts restrictions on $\boldsymbol{\beta}$. Solving this equation leads to

$$\boldsymbol{\beta}' = t(\boldsymbol{C}\mathbf{1})^- + \boldsymbol{\theta}(\boldsymbol{C}\mathbf{1})^{o'},$$

where $\boldsymbol{\theta}$ is an arbitrary vector of proper size. Moreover,

$$\boldsymbol{x}'\mathbf{1}^o = t(\boldsymbol{C}\mathbf{1})^-\boldsymbol{C}\mathbf{1}^o + \boldsymbol{\theta}(\boldsymbol{C}\mathbf{1})^{o'}\boldsymbol{C}\mathbf{1}^o + \widetilde{\boldsymbol{e}},$$

where $\widetilde{\boldsymbol{e}} \sim N_n(\mathbf{0}, \mathbf{1}^{o'}\boldsymbol{V}\mathbf{1}^o)$. In this model the MLE is obtained via

$$\widehat{\boldsymbol{\beta}}'\boldsymbol{C}\mathbf{1}^o = t(\boldsymbol{C}\mathbf{1})^-\boldsymbol{C}\mathbf{1}^o + \widehat{\boldsymbol{\theta}}(\boldsymbol{C}\mathbf{1})^{o'}\boldsymbol{C}\mathbf{1}^o,$$

where

$$\begin{aligned}\widehat{\boldsymbol{\theta}}(\boldsymbol{C}\mathbf{1})^{o'}\boldsymbol{C}\mathbf{1}^o &= (\boldsymbol{x}' - t(\boldsymbol{C}\mathbf{1})^-\boldsymbol{C})\mathbf{1}^o(\mathbf{1}^{o'}\boldsymbol{V}\mathbf{1}^o)^{-1}\mathbf{1}^{o'}\boldsymbol{C}'(\boldsymbol{C}\mathbf{1})^o \\ &\quad \times((\boldsymbol{C}\mathbf{1})^{o'}\boldsymbol{C}\mathbf{1}^o(\mathbf{1}^{o'}\boldsymbol{V}\mathbf{1}^o)^{-1}\mathbf{1}^{o'}\boldsymbol{C}'(\boldsymbol{C}\mathbf{1})^o)^-(\boldsymbol{C}\mathbf{1})^{o'}\boldsymbol{C}\mathbf{1}^o.\end{aligned}$$

from which $\widehat{\boldsymbol{\beta}}$ can be obtained under certain conditions on $\boldsymbol{C}$. □

In the example it was supposed that that $\boldsymbol{x}'\mathbf{1} = t$ which implied that $\mathbf{1}'\boldsymbol{V} = \mathbf{0}$. However, as noted above we may assume to have models where $\boldsymbol{V}$ is singular without any exact restrictions on $\boldsymbol{x}$. When restrictions are put on the covariance matrix we have restrictions on the random variable which only hold with probability 1. Therefore, it has in this case also to be assumed that data belongs to a proper subspace which indeed may be difficult to verify. Moreover, for the linear model

$$\boldsymbol{x}' = \boldsymbol{\beta}'\boldsymbol{C} + \boldsymbol{e}', \qquad \boldsymbol{e} = (\mathbf{0}, \sigma^2\boldsymbol{V})$$

with restrictions

$$\boldsymbol{\beta}'\boldsymbol{K} = \boldsymbol{h}$$

the situation can be described via the following model:

$$(\boldsymbol{x}' : \boldsymbol{h}) = \boldsymbol{\beta}'(\boldsymbol{C} : \boldsymbol{K}) + \boldsymbol{e}', \qquad \boldsymbol{e} = (\mathbf{0}, \sigma^2\boldsymbol{W}),$$

where

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{V} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

1.4 The General Multivariate Linear Model

We will study models which are based on an underlying multivariate normal distribution. Closely connected to linearity is the multivariate normal distribution since a linear function of a normal variable also is normally distributed. The theory around the normal distribution is well developed and one can among others show that the general linear model under certain conditions belongs to the exponential family which is very important. For example, there exist complete and sufficient statistics. Moreover all moments and cumulants are at our disposal.

The general linear multivariate model equals

$$\boldsymbol{X} = \boldsymbol{BC} + \boldsymbol{E}, \tag{1.5}$$

where $\boldsymbol{X} : p \times n$ is a random matrix which corresponds to the observations, $\boldsymbol{B} : p \times k$ is an unknown parameter matrix and $\boldsymbol{C} : k \times n$ is a known design matrix. Furthermore, $\boldsymbol{E} \sim N_{p,n}(\boldsymbol{0}, \boldsymbol{\Sigma}, \boldsymbol{I})$, where $\boldsymbol{\Sigma}$ is an unknown p.d. matrix. According to the model specifications the model consists of independently distributed columns and therefore $\boldsymbol{C}$ is also called a between individuals design matrix. In order to be able to draw any conclusions from the model we have to estimate the unknown parameters $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$. Following the statistical paradigm we also have to verify the model and this usually takes place with the help of residuals.

If looking at the likelihood function, $L(\boldsymbol{B}, \boldsymbol{\Sigma})$, we have

$$\begin{aligned} L(\boldsymbol{B}, \boldsymbol{\Sigma}) &\propto |\boldsymbol{\Sigma}|^{n/2} \exp(-1/2 \mathrm{tr}\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_0 - \boldsymbol{BC})()'\}) \\ &= |\boldsymbol{\Sigma}|^{n/2} \exp(-1/2 \mathrm{tr}\{\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_0 + \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_0\boldsymbol{P}_{C'} - \boldsymbol{BC})()'\}), \end{aligned}$$

where $\propto$ denotes proportional to and

$$\boldsymbol{S}_0 = \boldsymbol{X}_0(\boldsymbol{I} - \boldsymbol{P}_{C'})\boldsymbol{X}_0'.$$

Let $\boldsymbol{S}$ be as $\boldsymbol{S}_0$ but with $\boldsymbol{X}_0$ replaced by $\boldsymbol{X}$. From here it follows that the model belongs to the exponential family and that $\boldsymbol{X}\boldsymbol{P}_{C'}$ and $\boldsymbol{\Sigma}$ are sufficient statistics. It can be shown that the statistics also are complete. The MLEs for $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$ are obtained from

$$\begin{aligned} \widehat{\boldsymbol{B}}\boldsymbol{C} &= \boldsymbol{X}\boldsymbol{P}_{C'}, \\ n\widehat{\boldsymbol{\Sigma}} &= \boldsymbol{S} \end{aligned} \tag{1.6}$$

since (1.6) is a linear consistent equation system in $\boldsymbol{B}$ and the likelihood is always smaller or equal to $(2\pi)^{-pn/2}|n^{-1}\boldsymbol{S}_0|^{n/2}\exp(-\mathrm{np}/2)$, where the upper bound is obtained when inserting $\widehat{\boldsymbol{B}}\boldsymbol{C}$ and $\widehat{\boldsymbol{\Sigma}}$.

Example 1.3. *Many variables:* In environmental monitoring one may use many chemical biomarkers. For example, in Sweden, among others one follows calcium, magnesium, sodium, potassium, sulphate, cloride, flouride, nitrogen, phosphor, conductivity and other substances/properties in lakes spread over the whole country. Observations are taken many times over the year. Imagine that we would like to compare two regions a specific year. Then one may select 20 lakes from each region and the response variables would be the above mentioned chemical variables where for example an average over the summer months may be used. The model for the data with 10 response variables and 40 observations equally distributed over 2 lakes can be presented in the following way:

$$\boldsymbol{X} = \boldsymbol{BC} + \boldsymbol{E},$$

where $\boldsymbol{X}$: 10×40, $\boldsymbol{B}$: 10×2 consists of the mean parameters, $\boldsymbol{E} \sim N_{10,40}(\boldsymbol{0}, \boldsymbol{\Sigma}, \boldsymbol{I})$ where $\boldsymbol{\Sigma}$: 10×10 is the the unknown dispersion matrix, and

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{1}_{20}' & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{1}_{20}' \end{pmatrix}.$$

□

Example 1.4. *Repeated measurements with unstructured mean:* Another strategy for comparing regions than presented in Eaxmple 1.3 would be to focus on one of the chemical variables, for example nitrogen. Moreover, instead of averaging over the summer months as in Example 1.3 we may use the measurements from June, July and August. Thus we can set up the following model

$$\boldsymbol{X} = \boldsymbol{BC} + \boldsymbol{E},$$

where $\boldsymbol{X}$: 3×40, $\boldsymbol{B}$: 3×2 consists of the mean parameters, $\boldsymbol{E} \sim N_{3,40}(\boldsymbol{0}, \boldsymbol{\Sigma}, \boldsymbol{I})$ where $\boldsymbol{\Sigma}$: 3×3 is the unknown dispersion matrix, and the between individuals design matrix $\boldsymbol{C}$ is as in Example 1.3. □

There are two natural follow up questions concerning the models presented in Examples 1.3 and 1.4. The first is concerned with the repeated measurements for nitrogen over the summer months. It would be of interest to use a linear models for these measurements, in particular if we would include data from some more months. Then we would have a complete analogy with analysis of growth curve data but here, instead of growth, nitrogen over time is studied. The second question is if we can analyze all ten chemical variables over time. In this case we have an analogy with a spatio-temporal model setting. Here, instead of geographic spatial information we have been observing different chemical variables. Both these extensions are outside the general multivariate linear model setting. Under certain restrictions they can be analyzed with bilinear regression models since the mean structure instead of linear is bilinear. This implies, among others, that the models do not belong to the exponential family.

1.5 Bilinear Regression Models: An Introduction

Throughout, *BRM* is used as an acronym for Bilinear Regression Model. In the end of the previous section it was noted that even under normality assumptions we have very natural models which do not belong to the exponential family. It has been noted in the previous section that if the model has a linear mean structure the model belongs to the exponential family. In this section, among others, it will be shown that a bilinear mean structure will put the model outside the exponential family and instead will belong to the curved exponential family. Remember that if a matrix is pre- and post-multiplied by other matrices we perform a bilinear transformation. Often the mean structure $\boldsymbol{ABC}$ is considered, where the unknown parameter is given by $\boldsymbol{B}$.

Hence, we have a bilinear model

$$\boldsymbol{X} = \boldsymbol{ABC} + \boldsymbol{E}, \tag{1.7}$$

where $\boldsymbol{X}$: $p \times n$, the unknown mean parameter matrix $\boldsymbol{B}$: $q \times k$, the two design matrices $\boldsymbol{A}$: $p \times q$ and $\boldsymbol{C}$: $k \times n$, and the error matrix $\boldsymbol{E}$ build up the model. Moreover, let $\boldsymbol{E}$ be normally distributed with independent columns, with mean $\boldsymbol{0}$, and a covariance matrix $\boldsymbol{\Sigma}$ for the elements within each column. Therefore, the density function for $\boldsymbol{X}$ is proportional to

$$|\boldsymbol{\Sigma}|^{-1/2n} \exp(-1/2 \mathrm{tr}\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{ABC})(\boldsymbol{X} - \boldsymbol{ABC})'\})$$

and after some manipulations it may be shown that this model belongs to the curved exponential family. For example, it can can be seen by a reparameterization, i.e let $\boldsymbol{A} = \boldsymbol{\Gamma}\binom{\boldsymbol{I}}{\boldsymbol{0}}\boldsymbol{T}$ where $\boldsymbol{\Gamma}$ is orthogonal and $\boldsymbol{T}$ is a non-singular matrix. Moreover, let $\boldsymbol{\Theta} = \boldsymbol{TB}$, $\boldsymbol{\Psi} = \boldsymbol{\Gamma}'\boldsymbol{\Sigma}\boldsymbol{\Gamma}$ and

$$\boldsymbol{\Psi}^{-1} = \begin{pmatrix} \boldsymbol{\Psi}^{11} & \boldsymbol{\Psi}^{12} \\ \boldsymbol{\Psi}^{21} & \boldsymbol{\Psi}^{22} \end{pmatrix}.$$

Then, the density function is proportional to

$$|\boldsymbol{\Psi}|^{-1/2n}\exp(-1/2(\mathrm{tr}\{\boldsymbol{\Psi}^{-1}\boldsymbol{X}\boldsymbol{X}'\} - 2\mathrm{tr}\{\boldsymbol{\Psi}^{11}\boldsymbol{BCX}_1'\} + \mathrm{tr}\{\boldsymbol{\Psi}^{12}\boldsymbol{BCX}_2'\} + \mathrm{tr}\{\boldsymbol{ABCC}'\boldsymbol{B}'\boldsymbol{A}'\}))$$

which shows that the model belongs to the curved exponential family.

The above mentioned model is often termed Growth Curve model and was introduced by Potthoff & Roy (1964), although very similar models had been considered earlier. The $\boldsymbol{A}$ matrix is often refereed to as the within individuals design matrix and $\boldsymbol{C}$ as in (1.5) is called between individuals design matrix.

A natural extension of the BRM is the following "Sum of Profiles" model

$$\boldsymbol{X} = \sum_{i=1}^{m} \boldsymbol{A}_i\boldsymbol{B}_i\boldsymbol{C}_i + \boldsymbol{E},$$

where the sample matrix $\boldsymbol{X}$: $p \times n$, the mean parameter matrices $\boldsymbol{B}_i$: $q_i \times k_i$, the within individual design matrices equal $\boldsymbol{A}_i$: $p \times q_i$ and the between individual design matrices $\boldsymbol{C}_i$: $k_i \times n$ are such that

$$\mathcal{C}(\boldsymbol{C}_m') \subseteq \mathcal{C}(\boldsymbol{C}_{m-1}') \subseteq \cdots \subseteq \mathcal{C}(\boldsymbol{C}_1'). \tag{1.8}$$

Let $\boldsymbol{E}$ be normally distributed with independent columns, mean $\boldsymbol{0}$, and a covariance matrix $\boldsymbol{\Sigma}$ for the elements within each column. Observe that instead of (1.8) we may suppose

$$\mathcal{C}(\boldsymbol{A}_m) \subseteq \mathcal{C}(\boldsymbol{A}_{m-1}) \subseteq \cdots \subseteq \mathcal{C}(\boldsymbol{A}_1). \tag{1.9}$$

The model will be termed extended bilinear regression model and in order to distinguish between (1.8) and (1.9) as well as indicate m in the profile expression $EBRM_B^m$ and $EBRM_W^m$ will be used, where the subscripts B and W stand for between and within depending on if (1.8) or (1.9) is assumed to hold. The conditions in (1.8) or (1.9) are mathematically motivated since they lead to explicit MLEs. There is an analogy with the Behrens-Fisher problem, i.e. we want to compare two groups concerning equality in mean with the additional assumption that random variables corresponding to the observations from different groups have different variances, i.e.

$$\boldsymbol{x}' = \boldsymbol{\mu}'\boldsymbol{C} + \boldsymbol{e}',$$

where $\boldsymbol{\mu}' = (\boldsymbol{\mu}_1' : \boldsymbol{\mu}_2')$,

$$\boldsymbol{C} = \left(\mathbf{1}_{n_1}' \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix} : \mathbf{1}_{n_2}' \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix}\right),$$

and

$$\boldsymbol{e}' \sim N_n(\mathbf{0}, \begin{pmatrix} \sigma_1^2 \boldsymbol{I}_{n_1} & 0 \\ 0 & \sigma_2^2 \boldsymbol{I}_{n_2} \end{pmatrix}), \qquad n = n_1 + n_2.$$

To compare $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ will not give any precise answer about differences between groups, i.e. their distributions, unless σ_i^2 is taken into account. If (1.8) or (1.9) do not hold we have instead of a common mean and different variances, different means and a common dispersion. The situation is called seemingly unrelated regression (SUR) and has been extensively studied in a univariate setting. In the multivariate case it becomes more difficult to interpret results and there are reasons to avoid this type of models.

We conclude the section by giving some examples.

Example 1.5. *BRM; Liming data:* For many years there is a problem with acidification and to help lakes to recover one is liming them to stimulate the recovering process. Below we present a data set which includes 20 lakes from 2 regions were pH concentration has been measured at three different depths. Since pH is highest close to surface and thereafter diminish we may tray to model the concentration with a linear model. Data is presented in Table 1. The following matrices are involved in the analyses: $\boldsymbol{X}$ is the random matrix which corresponds to data and $\boldsymbol{X} \sim N_{3,20}(\boldsymbol{ABC}, \boldsymbol{\Sigma}, \boldsymbol{I})$, where $\boldsymbol{B}$ is an unknown parameter matrix and $\boldsymbol{\Sigma}$ is p.d. but unstructured,

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0.5 \\ 1 & 5 \\ 1 & 10 \end{pmatrix}, \qquad \boldsymbol{C} = \left(\mathbf{1}'_{10} \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix} : \mathbf{1}'_{10} \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right).$$

□

Example 1.6. *Melatonin and acute severe depression:* Already more than 20 years ago depression was studied concerning its relation to various hormones, in particular melatonin. Among others the melatonin peak level was found lowered in acutely ill depressed patients in comparison to healthy subjects. Melatonin measurements are illustrated in Figure 1. Peak levels remained low when these patients were re-examined during remission. Therefore melatonin levels may be viewed as a bio-marker for depression. Typical for melatonin as well as some other hormones (e.g. cortisol) is that they follow a day and night cycle.

Table 1: Selected data from the integrated monitoring of the effects of liming project at SLU, Sweden. pH = minus the decimal logarithm of the hydrogen ion activity.

lake	depth	pH	region	lake	depth	pH	region
1	0.5	6.72	1	11	0.5	7.29	2
1	5.0	6.61	1	11	5.0	6.78	2
1	10.0	6.41	1	11	10.0	6.76	2
2	0.5	6.80	1	12	0.5	6.91	2
2	5.0	6.80	1	12	5.0	6.91	2
2	10.0	6.70	1	12	10.0	6.71	2
3	0.5	7.16	1	13	0.5	7.23	2
3	5.0	7.12	1	13	5.0	7.37	2
3	10.0	7.01	1	13	10.0	7.10	2
4	0.5	7.17	1	14	0.5	6.81	2
4	5.0	7.20	1	14	5.0	6.68	2
4	10.0	7.08	1	14	10.0	6.18	2
5	0.5	6.96	1	15	0.5	6.66	2
5	5.0	6.68	1	15	5.0	6.47	2
5	10.0	6.48	1	15	10.0	6.17	2
6	0.5	7.23	1	16	0.5	6.89	2
6	5.0	7.02	1	16	5.0	6.59	2
6	10.0	6.80	1	16	10.0	6.19	2
7	0.5	6.87	1	17	0.5	6.98	2
7	5.0	6.73	1	17	5.0	6.64	2
7	10.0	6.43	1	17	10.0	6.24	2
8	0.5	7.15	1	18	0.5	6.88	2
8	5.0	7.18	1	18	5.0	7.01	2
8	10.0	6.80	1	18	10.0	6.71	2
9	0.5	7.23	1	19	0.5	7.01	2
9	5.0	7.03	1	19	5.0	6.90	2
9	10.0	6.73	1	19	10.0	6.80	2
10	0.5	7.24	1	20	0.5	7.20	2
10	5.0	7.19	1	20	5.0	7.17	2
10	10.0	6.99	1	20	10.0	7.07	2

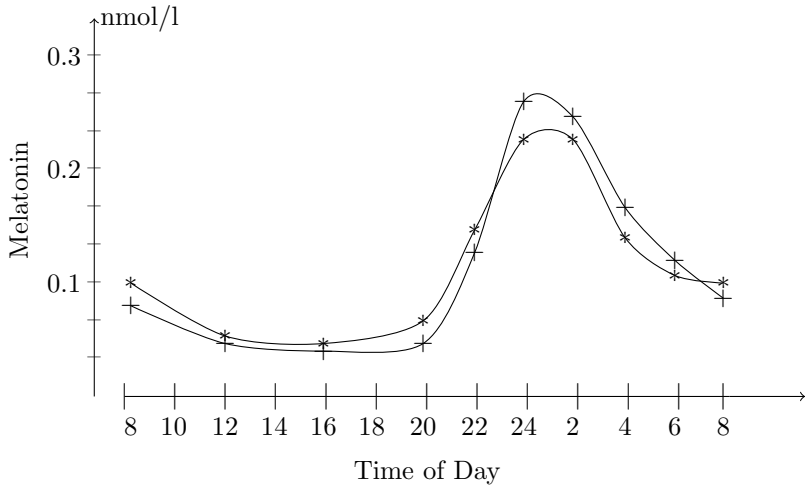

Figure 1: Serum melatonin for acute depressed patients (∗) and a control group of healthy individuals (+). Group averaged sample means have been joined in the figure.

The following model may be used: $\boldsymbol{X} \sim N_{10,60}(\boldsymbol{ABC}, \boldsymbol{\Sigma}, \boldsymbol{I})$, where ($\omega = \pi/24$)

$$\boldsymbol{A} = \begin{pmatrix} 1 & \sin(\omega) & \cos(\omega) & \sin(2\omega) & \cos(2\omega) \\ 1 & \sin(4\omega) & \cos(4\omega) & \sin(4*2\omega) & \cos(4*2\omega) \\ 1 & \sin(8\omega) & \cos(8\omega) & \sin(8*2\omega) & \cos(8*2\omega) \\ 1 & \sin(12\omega) & \cos(12\omega) & \sin(12*2\omega) & \cos(12*2\omega) \\ 1 & \sin(14\omega) & \cos(14\omega) & \sin(14*2\omega) & \cos(14*2\omega) \\ 1 & \sin(16\omega) & \cos(16\omega) & \sin(16*2\omega) & \cos(16*2\omega) \\ 1 & \sin(18\omega) & \cos(18\omega) & \sin(18*2\omega) & \cos(18*2\omega) \\ 1 & \sin(20\omega) & \cos(20\omega) & \sin(20*2\omega) & \cos(20*2\omega) \\ 1 & \sin(22\omega) & \cos(22\omega) & \sin(22*2\omega) & \cos(22*2\omega) \\ 1 & \sin(24\omega) & \cos(24\omega) & \sin(24*2\omega) & \cos(24*2\omega) \end{pmatrix},$$

$$\boldsymbol{C} = \left(\mathbf{1}'_{28} \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix} : \mathbf{1}'_{32} \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right).$$

□

Example 1.7. *BRM, Potthoff & Roy (1964) classical data set.* Data consist of growth measurements, i.e. the distance in mm. from the center of the pituitary to the pterygo-maxillary fissure, for 11 girls and 16 boys at ages $t_1 = 8$, $t_2 = 10$, $t_3 = 12$, and $t_4 = 14$. The design matrices equal

$$\boldsymbol{A} = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ 1 & t_3 \\ 1 & t_4 \end{pmatrix}, \quad \text{Linear growth,} \quad \boldsymbol{A} = \begin{pmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ 1 & t_3 & t_3^2 \\ 1 & t_4 & t_4^2 \end{pmatrix}, \quad \text{Quadratic growth,}$$

$$\boldsymbol{C} = \left(\mathbf{1}'_{11} \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix} : \mathbf{1}'_{16} \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right).$$

Then, the model is given by either $\boldsymbol{X} \sim N_{4,27}(\boldsymbol{A}_1\boldsymbol{B}\boldsymbol{C}, \boldsymbol{\Sigma}, \boldsymbol{I})$ or $\boldsymbol{X} \sim N_{4,27}(\boldsymbol{A}_2\boldsymbol{B}\boldsymbol{C}, \boldsymbol{\Sigma}, \boldsymbol{I})$. The data is presented in Table 2 and illustrated in Figure 2. One can see that there is a difference between boys and girls and later we are going to investigate if there is a model which can be analyzed statistically, including a validation of the model, where the difference between gender can be tested.

Table 2: Four repeated measurements were taken at ages $t_1 = 8$, $t_2 = 10$, $t_3 = 12$, and $t_4 = 14$ from 11 girls and 16 boys.

id	gender	t_1	t_2	t_3	t_4	id	gender	t_1	t_2	t_3	t_4
1	F	21.0	20.0	21.5	23.0	14	M	23.0	22.5	24.0	27.5
2	F	21.0	21.5	24.0	25.5	15	M	25.5	27.5	26.5	27.0
3	F	20.5	24.0	24.5	26.0	16	M	20.0	23.5	22.5	26.0
4	F	23.5	24.5	25.0	26.5	17	M	24.5	25.5	27.0	28.5
5	F	21.5	23.0	22.5	23.5	18	M	22.0	22.0	24.5	26.5
6	F	20.0	21.0	21.0	22.5	19	M	24.0	21.5	24.5	25.5
7	F	21.5	22.5	23.0	25.0	20	M	23.0	20.5	31.0	26.0
8	F	23.0	23.0	23.5	24.0	21	M	27.5	28.0	31.0	31.5
9	F	20.0	21.0	22.0	21.5	22	M	23.0	23.0	23.5	25.0
10	F	16.5	19.0	19.0	19.5	23	M	21.5	23.5	24.0	28.0
11	F	24.5	25.0	28.0	28.0	24	M	17.0	24.5	26.0	29.5
						25	M	22.5	25.5	25.5	26.0
						26	M	23.0	24.5	26.0	30.0
						27	M	22.0	21.5	23.5	25.0

□

Example 1.8. *$EBRM^3_B$:* Let us start from the very beginning and suppose that we have a random vector $\boldsymbol{x}$ associated to observations which follow the model

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{e},$$

where $\boldsymbol{e} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$. Now assume that there exist a linear relation among the components in $\boldsymbol{\mu}$, i.e. $\boldsymbol{\mu} \in \mathcal{C}(\boldsymbol{A})$. Thus, $\boldsymbol{\mu} = \boldsymbol{A}\boldsymbol{\beta}$ for some $\boldsymbol{\beta}$ and $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{e}$. Moreover, suppose that we have n independent observations which all have the same within individual model $\boldsymbol{\mu} \in \mathcal{C}(\boldsymbol{A})$ and suppose that there additionally exists a linear model between the independent observations. For example, there are three groups of individuals; one corresponding to a placebo treatment and the others corresponding to two different treatments, respectively. Thus we end up in the following model

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{B}\boldsymbol{C} + \boldsymbol{E},$$

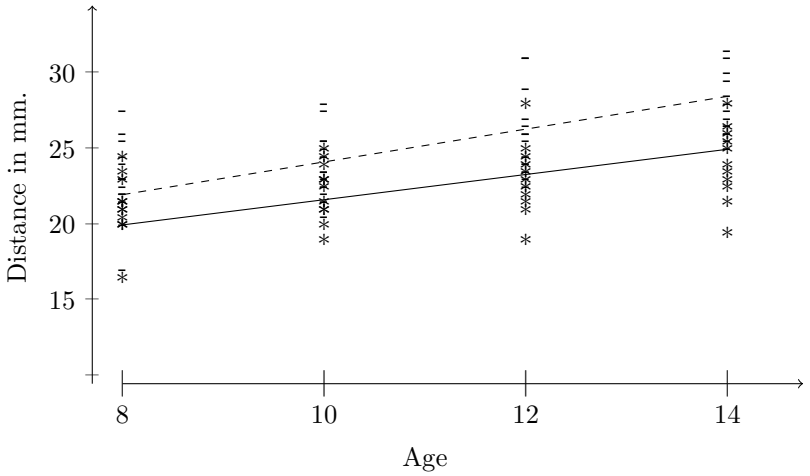

Figure 2: The distance in mm. from the center of the pituitary to the pterygomaxillary fissure in boys (–) and girls (- -) at ages 8,10,12 and 14.

where $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$, $\boldsymbol{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$, $\boldsymbol{E} \sim N_{n,p}(\boldsymbol{0}, \boldsymbol{I}, \boldsymbol{\Sigma})$ and

$$\boldsymbol{C} = \begin{pmatrix} 1 & 1 & \ldots & 1 & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & 1 & 1 & \ldots & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & 1 & 1 & \ldots & 1 \end{pmatrix}.$$

Furthermore, assume that we have a polynomial growth. Then the Vandermonde matrix, for example,

$$\boldsymbol{A} = \begin{pmatrix} 1 & t_1 & \ldots & t_1^{q-1} \\ 1 & t_2 & \ldots & t_2^{q-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_p & \ldots & t_p^{q-1} \end{pmatrix}$$

describes the connection between growth and time. In this model all individuals follow the same polynomial growth model. However, if each treatment group follows a polynomial of different order we may for example have the following model

$$\boldsymbol{X} = \boldsymbol{A}_1\boldsymbol{B}_1\boldsymbol{C}_1 + \boldsymbol{A}_2\boldsymbol{B}_2\boldsymbol{C}_2 + \boldsymbol{A}_3\boldsymbol{B}_3\boldsymbol{C}_3 + \boldsymbol{E},$$

where

$$\boldsymbol{C}_1 = \begin{pmatrix} 1 & 1 & \ldots & 1 & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & 1 & 1 & \ldots & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & 1 & 1 & \ldots & 1 \end{pmatrix},$$

$$\boldsymbol{C}_2 = \begin{pmatrix} 1 & 1 & \ldots & 1 & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & 1 & 1 & \ldots & 1 & 0 & 0 & \ldots & 0 \end{pmatrix},$$

$$\boldsymbol{C}_3 = \begin{pmatrix} 1 & 1 & \dots & 1 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \end{pmatrix},$$

$$\boldsymbol{A}_1 = \begin{pmatrix} 1 & t_1 & \dots & t_1^{q-3} \\ 1 & t_2 & \dots & t_2^{q-3} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_p & \dots & t_p^{q-3} \end{pmatrix}, \qquad \boldsymbol{B}_1 = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3),$$

$$\boldsymbol{A}_2' = \begin{pmatrix} t_1^{q-2} & t_2^{q-2} & \dots & t_p^{q-2} \end{pmatrix}, \qquad \boldsymbol{B}_2 = (\beta_3, \beta_4),$$

$$\boldsymbol{A}_3' = \begin{pmatrix} t_1^{q-1} & t_2^{q-1} & \dots & t_p^{q-1} \end{pmatrix}, \qquad \boldsymbol{B}_3 = \beta_5.$$

Observe that $\mathcal{C}(\boldsymbol{C}_3') \subseteq \mathcal{C}(\boldsymbol{C}_2') \subseteq \mathcal{C}(\boldsymbol{C}_1')$. The above example implies, for example, that the mean of the placebo group and the treatment groups respectively equal

$$\begin{aligned} &\beta_{11} + \beta_{12}t + \cdots + \beta_{1(q-2)}t^{q-3}, \\ &\beta_{21} + \beta_{22}t + \cdots + \beta_{2(q-2)}t^{q-3} + \beta_{2(q-1)}t^{q-2}, \\ &\beta_{31} + \beta_{32}t + \cdots + \beta_{3(q-2)}t^{q-3} + \beta_{3(q-1)}t^{q-2} + \beta_{3q}t^{q-1}. \end{aligned}$$

□

Example 1.9. $EBRM_W^3$ Now the model in Example 1.8 will be reconsidered. The example indicates how the $EBRM_B^3$ and $EBRM_W^3$ are related. However, in general the relation between $EBRM_B^3$ and $EBRM_W^3$ is not so clear. It follows that the following model is equivalent to the model in Example 1.8:

$$\boldsymbol{X} = \boldsymbol{A}_1\boldsymbol{\Theta}_1\boldsymbol{C}_1 + \boldsymbol{A}_2\boldsymbol{\Theta}_2\boldsymbol{C}_2 + \boldsymbol{A}_3\boldsymbol{\Theta}_3\boldsymbol{C}_3 + \boldsymbol{E},$$

where

$$\begin{aligned} \boldsymbol{C}_1 &= \begin{pmatrix} 0 & 0 & \dots & 0 & 1 & 1 & \dots & 1 & 0 & 0 & \dots & 0 \end{pmatrix}, \\ \boldsymbol{C}_2 &= \begin{pmatrix} 0 & 0 & \dots & 0 & 1 & 1 & \dots & 1 & 0 & 0 & \dots & 0 \end{pmatrix}, \\ \boldsymbol{C}_3 &= \begin{pmatrix} 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 1 & 1 & \dots & 1 \end{pmatrix}, \end{aligned}$$

$$\boldsymbol{A}_1 = \begin{pmatrix} 1 & t_1 & \dots & t_1^{q-1} \\ 1 & t_2 & \dots & t_2^{q-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_p & \dots & t_p^{q-1} \end{pmatrix}, \qquad \boldsymbol{\Theta}_1 = (\boldsymbol{\beta}_1', \beta_3, \beta_5)',$$

$$\boldsymbol{A}_2 = \begin{pmatrix} 1 & t_1 & \dots & t_1^{q-2} \\ 1 & t_2 & \dots & t_2^{q-2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_p & \dots & t_p^{q-2} \end{pmatrix}, \qquad \boldsymbol{\Theta}_2 = (\boldsymbol{\beta}_2', \beta_4)',$$

$$\boldsymbol{A}_3 = \begin{pmatrix} 1 & t_1 & \dots & t_1^{q-3} \\ 1 & t_2 & \dots & t_2^{q-3} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_p & \dots & t_p^{q-3} \end{pmatrix}, \qquad \boldsymbol{\Theta}_3 = \boldsymbol{\beta}_3.$$

The interesting point is that now $\mathcal{C}(\boldsymbol{A}_3) \subseteq \mathcal{C}(\boldsymbol{A}_2) \subseteq \mathcal{C}(\boldsymbol{A}_1)$ holds instead of $\mathcal{C}(\boldsymbol{C}_3') \subseteq \mathcal{C}(\boldsymbol{C}_2') \subseteq \mathcal{C}(\boldsymbol{C}_1')$. Moreover, when considering $EBRM_B^3$ $\{\boldsymbol{B}_i\}$, $i = 1, 2, 3$, are the objects of interest whereas in $EBRM_W^3$ the parameters $\{\boldsymbol{\Theta}_i\}$, $i = 1, 2, 3$, are of interest. For example, if estimability conditions are considered it is not necessary that the estimability of $\boldsymbol{B}_1$ implies the estimability of $\boldsymbol{\Theta}_1$. Of course if $\{\boldsymbol{B}_i\}$ is estimable then also $\{\boldsymbol{\Theta}_i\}$, $i = 1, 2, 3$, but usually we are not interested to estimate all parameters in $\{\boldsymbol{B}_i\}$, $i = 1, 2, 3$, uniquely and then it is not so easy to find out estimability conditions for $\boldsymbol{\Theta}_i$. Moreover, to derive $D[\boldsymbol{\Theta}_i]$ from $D[\boldsymbol{B}_i]$ without knowledge about $C[\boldsymbol{B}_i, \boldsymbol{B}_j]$, $i \neq j$, is impossible. □