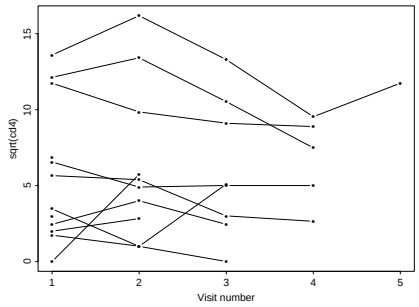
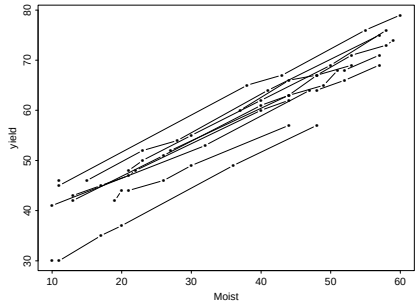
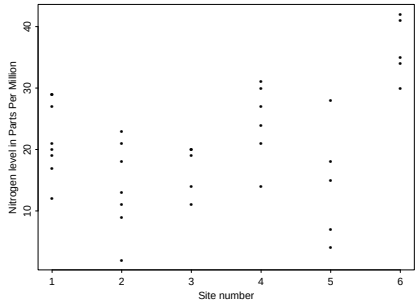
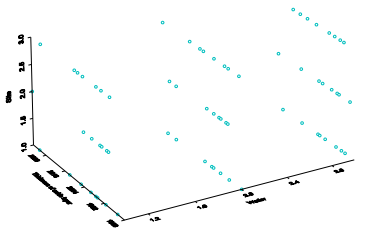
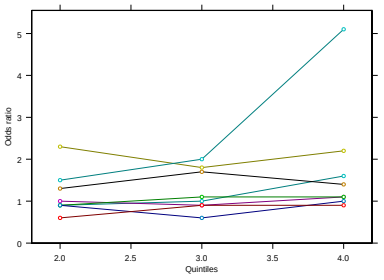










University of Oslo
Department of Mathematics

A simulation study of methods for estimation and testing in mixed effects models

Harald Fekjær

Cand.scient. thesis

1999

About the cover:

Figures on the cover page show different examples of data suitable for mixed effects models. The data are taken from different fields where mixed effects models are in practical use. Labeled from top of page these are:

1. Medicine:

   Repeated CD4 measurements for patients in a clinical AIDS trial. Data from example 10.10 in SAS system for mixed effects models (Littell, Milliken, Stroup & Wolfinger 1996). Plot are only for clink number 5.

2. Agriculture:

   Measurement of yield in ten types of winter wheat, under different amounts of preplant moisture in the soil. Data from example 7.2.1 in SAS system for mixed effects models (Littell et al. 1996).

3. Environment.

   Repeated measurement of nitrogen levels in the Mississippi river. Data are from example 4.2 in SAS system for mixed effects models (Littell et al. 1996).

4. Economics

   Demand per capita in 13 different US states over time. Data are from example 3.4 in SAS system for mixed effects models (Littell et al. 1996).

5. Technology:

   Variation of thickness of silicon on wafers in semi conductor production (for different sites and wafers). Example from 4.4 in SAS system for mixed effects models (Littell et al. 1996).

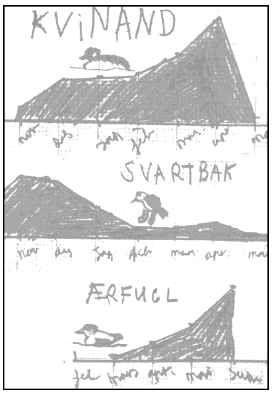
6. Nutrition / Medicine:

   Relative risk of breast cancer for different cholesterol levels. Data are from Rønnevik (1999).

Preface

> The business of the statistician is to catalyze the scientific learning process.
> George E. P. Box (1919-)

From an early age, I have always been interested in research and man's quest for better knowledge about the world around him. My first introduction to statistics came when I was at the age of twelve, and wanted to learn more about the variations of seabird populations around my parents house at Nesodden (Norway). I made a design where I counted the seabirds in a specific area 3-7 times a week. With this design, I counted birds for a little over one year, and my parents taught me about mean, median and some other statistical methods suitable for drawing conclusions from my material. With my interest in nature, everybody thought I was going to become a biologist, but fate wanted it another way. And looking back one may also see some signs of a coming statistician.



"My first statistics"

Several years later, I attended a statistical course at the University of Oslo, and soon realized that this was an ideal way for me to combine my interest in applied mathematics, information technology and general reseach.

Later, while I was teaching at the Section of Medical Statistics, I was introduced to Petter Laake, who showed me his ideas which later became the basic for this thesis. Since then, I have participated in several projects and teaching assignments at the Section of Medical Statistics, and I am now looking forward to continued work here in an expanding group with superb colleagues and work environment.

I wish to thank Petter Laake, Inge Helland, Bjørn Møller, Nils Lid Hjort, Thore Egeland and Kåre Osnes for many helpful remarks and considerations. Petter Laake has been the best supervisor I could get, and it is with great pleasure I now look back on my first glimpse of the statistical research field. And last but not least, I also wish to send a special thanks to Harald Weedon, my family and my girlfriend Susanne for support and help with my written English.

Harald Fekjær, University of Oslo, June 15, 1999

Contents

Aim of study

> Mathematics seems to endow one with something like a new sense.
> Charles Darwin (1809-1882)

During the last decade, the mixed effects model has become one of the most used statistical models in medical research. Traditionally, estimation in mixed effects model has been done using different kinds of quadratic sums methods, but with increased computing power and new software, methods like restricted Maximum Likelihood (ML) and Gibbs sampling have gained popularity.

This raises several questions about the efficiency of the different estimation methods, both generally and relative to each other. For some estimation methods we have analytical solutions for the distributions, but with unbalanced designs, analytic solutions become very hard or impossible to find. This means that we have to rely on simulation studies. One of the first such simulation studies was published by Swallow in 1981 (Swallow 1981), followed up with Swallow & Monahan (1984). Later, there have been many articles on other special cases, which often where based on a special application, but to my knowledge no large studies have been published to this date.

Swallow & Monahan (1984) is the classical article which contains many interesting results, but some questions still remain. One example of this is the performance of the Gibbs sampler. At the time of Swallow & Monahan's article this estimator was almost in no practical use, but it has since then become quite popular. Other central questions arise around the different testing methods used in mixed effects models. Today the Wald test is implemented in several statistical packages, but is this a good choice? (e.g. Does perform reasonable well, as it does in most fixed effects cases?). The aim of this study is to shed some light on these questions, and to come up with practical advice for uses of estimation and testing methods in mixed effects models.

Chapter 1

Introduction to random and mixed effects models

> All models are wrong, but some are useful
> George E. P. Box (1919-)

From elementary statistical courses, the standard regression model is well known:

$$Y = X\beta + \varepsilon$$

where:

- $X = X_{mxp}$ is the design (covariate) matrix.
- $\beta = \beta_{p\times 1}$ is the parameter vector.
- $\varepsilon = \varepsilon_{m\times 1}$ is the error (rest) term,
  with $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_m)$ and $\varepsilon_i \sim N(0, \sigma_e^2)$.
  (E.g. independent normal distributed error terms.)

As we can see, the model is based on independent observations. This is a relevant model for some simple designs, but in many practical cases we have data that do not fulfill this assumption. One classical example of this is repeated measurements, where we in most cases will have an individual effect that creates dependent observations.

One solution to this problem could be to include the individual effect in the parameters. This is done by giving each individual its own parameters,

something that in practice leads to very many parameters. This can sometimes work for data with many replications, but in most cases we soon run into problems of overparametrization. These problems arise as we estimate a large number of parameters from a relatively small dataset. The added parameters increase the uncertainty in the model, and with few repetitions for each individual the model will often "collapse" without giving any good estimates for the parameters.

As overparametrization leaves the standard method useless in many practical cases, we must look for other methods that better include dependencies from designs such as repeated measurements. In this chapter, we will give a brief introduction to the mixed effects models, that solves this problem by adding so called random effects to the standard regression model.

1.1 Mixed effect models; Theory and examples

Statistics often involve dealing with repeated measurement data. It could be data from various fields such as medicine, economics and biology. This could typically be samples over time for each individual. One example of this is plotted in figure 1.1, where we find repeated CD4 measurements for patients in a clinical trial for AIDS patients[1].

Since data can be modeled in many different ways, we have to make some assumptions regarding the underlying model to perform an analysis. Maybe the most obvious model would be to assume that each person has his/hers own basic linear curve, with some noise around it. If we now assume that both this noise, the individual intercepts and the individual slopes, all are normally (Gaussian) distributed, we encounter a normal linear mixed effects model. If we now have m observations, this model can be written in the following form for each individual:

$$Y = X\beta + Z\alpha + \varepsilon$$

where:

- $X = X_{m\times p}$ is the design (covariate) matrix of the fixed effects.
- $\beta = \beta_{p\times 1}$ is the parameter vector for the fixed effects.

[1]Data are from example 10.10 in SAS system for mixed effects models (Littell et al. 1996).

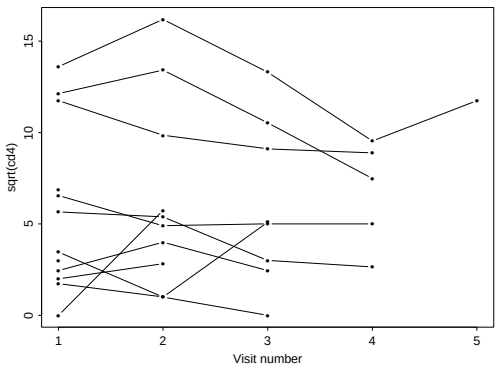


Figure 1.1: Example of repeated measurement data; CD4 levels from a clinical trial for AIDS patients (plotted only for clinck number 5).

- $Z = Z_{m \times q}$ is the design (covariate) matrix for the random effects.
- $\alpha = \alpha_{q \times 1}$ is the vector of random effects effects with $\alpha = (\alpha_1, \ldots, \alpha_q)$ and $\alpha_i \sim N(0, \sigma^2_{\alpha_i})$.
- $\varepsilon = \varepsilon_{m \times 1}$ is the error term, with the variation of the observation (as sampling error etc.). It can be written as $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_m)$, with $\varepsilon_i \sim N(0, \sigma^2_\varepsilon)$.

Remarks:

- We can assume that α has zero mean, because any mean can be included in the fixed effects.
- We can also write the model as: $N(x\beta, \Sigma)$, where Σ is the variance-covariance matrix of both the random effects and the error term.
- This model is written for one individual, but it is also possible to express a general formula for all individuals. In this case, we must repeat some of the random effects terms several times.

In the clinical AIDS trial, X will be the covariates connected to the visits, β will be the parameters for the overall population curve and ε the measuring error and stochastic variation around the model. This is the usual linear model we have seen in the introductory courses in statistics, but because we have repeated measurements we must extend the model somewhat. We

therefore introduce the term $Z\alpha$ for the individual effects. Here Z has two columns (Z_1, Z_2), where Z_1 is the intercept and Z_2 the slope. In the next chapter, we will see how we for this model can make some estimates of both the population slope, the individual variation and the variation of each observation. This could typically be used for observing effects of two different treatments (see figure 1.2).

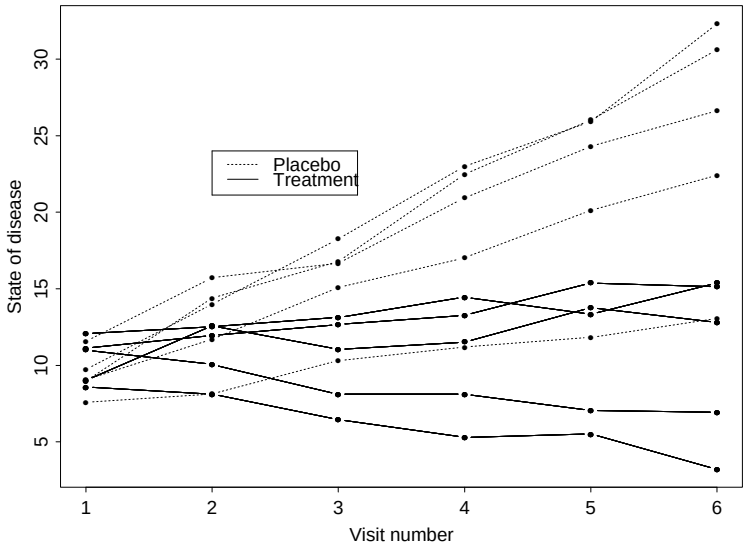


Figure 1.2: Typical repeated measurements data from a clinical trial.

Another example is the measurement of yield in ten types of winter wheat under different amounts of preplant moisture in the soil, shown in figure 1.3[2]. In this example we see that the linear structure is much more distinct, with a clear linear trend.

In addition to these linear models, we could also have non-linear structures as seen in figure 1.4. These data are from a biochemical experiment measuring velocity in cells with and without treatment by Puromycin[3], and here we see that the curves increase to a point, for then to slowly stabilize.

[2]Data from example 7.2.1 in SAS system for mixed effects models (Littell et al. 1996).

[3]Example from S-PLUS 4 manual "Guide to Statistics" (Data Analysis Products Division MathSoft Inc. 1997).

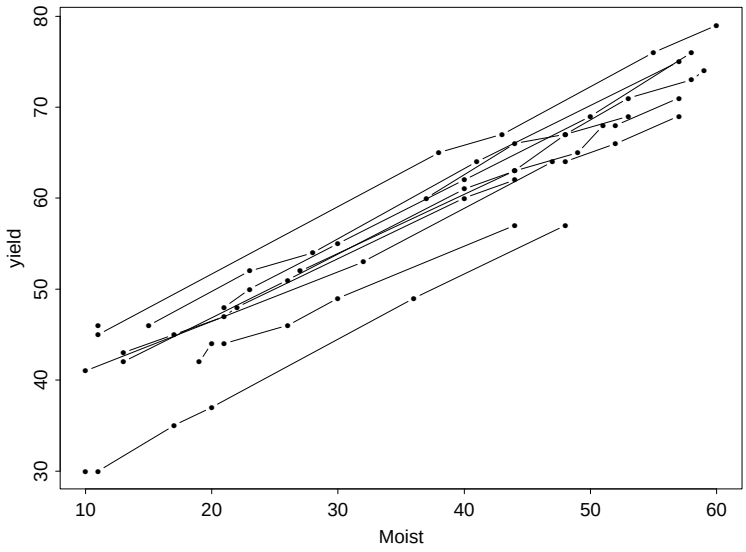


Figure 1.3: Example of mixed effects data with clear linear structure; Yield in ten types of winter wheat.

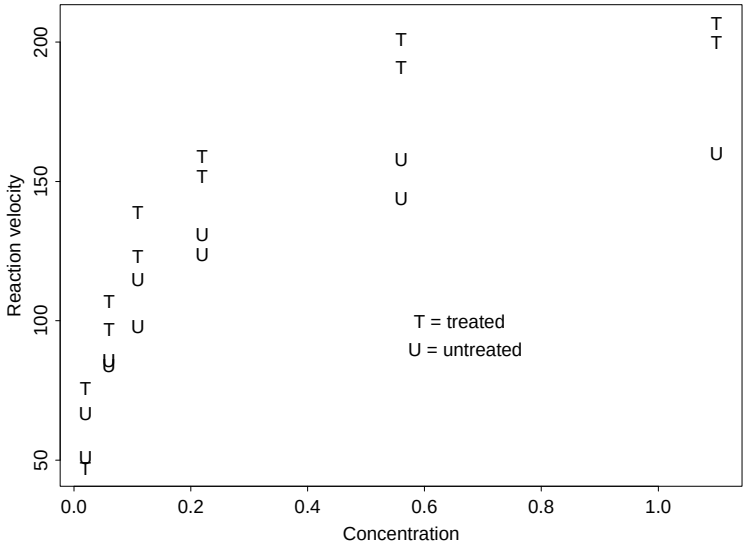


Figure 1.4: Example of data suitable for a non linear mixed effects model; Velocity in cells with and without treatment by Puromycin.

21	21	20	14	7	41
27	11	19	24	15	42
29	18	20	30	18	35
17	9	11	21	4	34
19	13	14	31	28	30
12	23	27	27	.	.
29	2	.	.	.	.
20	.	.	.	.	.
20	.	.	.	.	.

Table 1.1: Data suitable for a random component model; Nitrogen measurements along the Mississippi River.

The mixed effects model could also be generalized with other distributions than the normal (gaussian). Some common examples of this include binomial and logistic distributions. All in all, the mixed effect model is a superset of generalized linear models (GLM), with its wide set of different distributions.

Good sources for learning more about mixed models, are: "Tutorial in Biostatistics; Using the General Linear Mixed Model to analyse unbalanced repeated measurs and longitudinal Data" (Chaan, Laird & Slasor 1997), "Methods and Applications of Linear Models" (Hocking 1996) and "SAS system for Mixed Models" (Littell et al. 1996).

1.2 One special case: The random effects model

One common special case of the mixed effects model, is the random effects model. In this model the fixed effect is just a constant, and the mixed effects have no curve or other pattern. This model makes the estimation somewhat easier, and frequently occurs in nature.

One example of this is the repeated measurement of nitrogen levels from the Mississippi river[4]. In this example we have repeated measurements from several places along the river, and wonder both about the variation within and between the different places. The data are shown in table 1.1, and plotted in figure 1.5.

[4]Data are from example 4.2 in SAS system for mixed effects models (Littell et al. 1996).

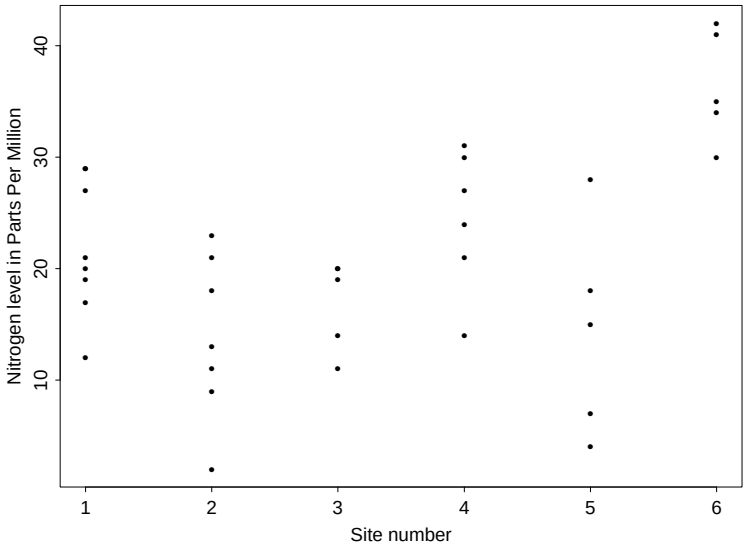


Figure 1.5: Example of a random component model; Nitrogen measurements from the Mississippi River.

As the data are unbalanced with just one random effect, this model is called a one-way unbalanced random effects model, and can be written as:

$$Y = \mu + \alpha_i + \varepsilon_{ij}$$

where:

- μ is a fixed constant
- α is the random effect with $\alpha_i \sim N(0, \sigma_a^2)$ for $i = 1, ..., m$
- ε_{ij} is the variation of the observation (as sampling error etc.) with $\varepsilon_{ij} \sim N(0, \sigma_e^2)$ for $i = 1, ..., m$ and $j = 1, ..., n_i$

Later, we shall use this model in our study of different estimating and testing methods. This model is chosen as it is both simple, perspicuous and in widespread use, while it still properly can give us some hints of what happens in the more complex cases of mixed effect models.

Chapter 2

Estimation and testing in mixed effects models

> Everything should be made as simple as possible, but not simpler.
> Albert Einstein (1879-1955)

In the previous chapter we have seen the model for mixed effects. This model can serve as a setting for the analysis, but for any practical use of the model we will need methods for estimating the unknown parameters. Estimation in mixed effects models has traditionally been done by different kinds of quadratic sums methods, but the development in computer power has opened up new possibilities such as likelihood based approaches and Gibbs sampling. In this chapter we will give a brief introduction to the most common estimation and testing methods used in mixed effects models.

2.1 Estimation methods for random and mixed effects

2.1.1 Analysis of variance (ANOVA) based estimation

The traditional method of estimating random components is through the quadratic forms of the ANOVA table. The basic method is to equal the observed sum of squares to their expectation, and solve the equations to find the estimates. One example of this for the balanced one-way random effects model, is shown below:

Source of variation	Degrees of freedom	Sum of squares	Expected mean sum of squares
Between	$m-1$	$\sum_{i=1}^{m}(\overline{y}_{i.}-\overline{y}_{..})^2$	$\sigma_a^2+\sigma_e^2/n$
Within	$mn-m$	$\sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij}-\overline{y}_{i.})^2$	σ_e^2

Table 2.1: Sum of squares for balanced random effects model.

$$Y=\mu+\alpha_i+\varepsilon_{ij}$$

where:

- μ is a fixed constant.
- α_i is the random effect with $\alpha_i \sim N(0,\sigma_a^2)$.
- ε_{ij} is the variation of the observation (as sampling error etc.) with $\varepsilon_{ij} \sim N(0,\sigma_e^2)$.
- $i=1,\ldots,m$ and $j=1,\ldots,n$

To find the ANOVA estimate for σ_e^2, we first equal the within sum of squares (see table 2.1) to its expectation:

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i.})^2=(mn-m)\hat{\sigma}_e^2.$$

And then solve this equation with regards to $\hat{\sigma}_e^2$, and get

$$\hat{\sigma}_e^2=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i.})^2}{mn-m}.$$

As for σ_e^2, we then find the σ_a^2 ANOVA estimate by using the sum of squares, but this time we start with the between sum of squares (see table 2.1):

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\overline{y}_{i.}-\overline{y}_{..})^2=(m-1)(\hat{\sigma}_a^2+\hat{\sigma}_e^2/n)$$

Then we solve this equation with regards to $\hat{\sigma}_a^2$, and get

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\overline{y}_{i.}-\overline{y}_{..})^2=(m-1)(\hat{\sigma}_a^2+\hat{\sigma}_e^2/n)\Leftrightarrow$$

$$\hat{\sigma}_a^2+\hat{\sigma}_e^2/n=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\overline{y}_{i.}-\overline{y}_{..})^2}{m-1}\Leftrightarrow$$

$$\hat{\sigma}_a^2=\frac{\sum_{i=1}^{m}(\overline{y}_{i.}-\overline{y}_{..})^2}{(m-1)}-\frac{\hat{\sigma}_e^2}{n}.$$

This basic method works well for balanced designs. Under normality assumptions it gives us the best quadratic unbiased estimator (e.g. the unbiased quadratic based estimator with the smallest variance), and is in almost all cases easy to compute. In addition it does not depend on any normal distribution assumptions for the basic estimating theory, and gives unbiased estimators even for non normal distributed data.

ANOVA for unbalanced designs

However, for unbalanced designs the basic ANOVA method does not give a unique solution. To address this problem, Henderson in 1953 described three different methods of generalizing the ANOVA approach to deal with unbalanced models. These methods have been named Henderson Method I, II and III or estimation based on Type I, II and III quadratic forms. They differ only by the quadratics uses in the estimation process. Below we will give a short introduction to the these three methods:

Henderson Type I

Henderson Type I uses quadratic forms that are analog to the one used in standard analysis of variance. For the unbalanced random effects model with $j=1,....,n_i$, these quadratic forms are found in table 2.2.

To find the ANOVA estimate for σ_e^2, we again equal the within sum of squares to its expectation, which gives

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i.})^2=(\sum_{i=1}^{m}n_i-m)\hat{\sigma}_e^2.$$

Then we solve the equation with regards to $\hat{\sigma}_e^2$, to get the estimate for the error term and get

$$\hat{\sigma}_e^2=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i.})^2}{\sum_{i=1}^{m}n_i-m}.$$

Source of variation	Sum of squares	Expected sum of squares
Between	$\sum_{i=1}^{m} n_i(\overline{y}_{i.} - \overline{y}_{..})^2$	$(N - \sum_{i=1}^{m} /N)\sigma_a^2 + (m-1)\sigma_e^2$
Within	$\sum_{i=1}^{m} \sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i.})^2$	$(\sum_{i=1}^{m} n_i - m)\sigma_e^2$

Table 2.2: Sum of squares for unbalanced random effects model. N is the total number of observations (E.g. $N = \sum_{i=1}^{m} n_i$).

As for the estimation of σ_a^2, we use the same method as for the balanced case, and set the between sum of squares equal to its expectation. Using this we get the following equation:

$$\sum_{i=1}^{m} n_i(\overline{y}_{i.} - \overline{y}_{..})^2 = (N - \sum_{i=1}^{m} /N)\hat{\sigma}_a^2 + (m-1)\hat{\sigma}_e^2$$

When solving this equation with regards to $\hat{\sigma}_a^2$, we get

$$\hat{\sigma}_a^2 = \frac{\sum_{i=1}^{m} n_i(\overline{y}_{i.} - \overline{y}_{..})^2 - (m-1)\sigma_e^2}{N - \sum_{i=1}^{m} n_i^2/N}.$$

In this example the quadratic forms are sum of squares, but in some other cases Hederson Method I gives us quadratic forms which are not non-negative definite, and hence also not sum of squares.

As the ANOVA estimator for balanced designs, Hederson Method I is usually quite easy to compute. It works fine in variance component models, but can not be used in mixed effects models. The model can in some ways be adapted to mixed effects (by treating the fixed effects as random etc.), but these approaches give biased estimats.

Henderson Method II:

Henderson Method II extends Method I to mixed effects model. In conduction Method II, we first treat the random effects as fixed, and then estimate the parameters by least squares. We then use these parameters to "remove" the fixed effect, by subtracting the estimated fixed effect term. In the end we use Method I on the "corrected" data.

Using the notation from section 1.1, this means treating $X\beta + Z\alpha$ as fixed effects, estimating $\hat{\beta}$ through least squares, and using Henderson Method I at $Y - X\hat{\beta}$.

This method works well on mixed effects data with no interaction between fixed and random factors, but with interaction between terms, this model gives biased estimates. Henderson method II is usually fairly easy to compute, but no analytic formulas for sampling variances are known.

Henderson Method III:

Henderson Method III uses quite a different approach than the first two methods. Instead of using analogs to the sum of squares from standard analysis of variance, it rather uses conventional least squares analysis of non-orthogonal data. In practice this means fitting an overparameterized model and sub-models, and using the rest terms as quadratic forms for the estimation process. One example of this is illustrated in Searle (1987).

Henderson Method III gives unbiased estimates even in mixed effects models with interaction between fixed and random factors, but there is no unique solution for the Henderson Method III estimate. This because there are too many sums of squares available, and there are no rules for which set of sums of squares to use. There is also no analytic proof that its sums of squares have any optimal properties for estimating variance components, but the method has turned out to be quite useful in practical applications. The method is also somewhat more demanding to compute than method I and II, but with today's fast computers this is usually no problem.

In our simple example of a one-way unbalanced random effects model, this approach gives different formulas for estimation. Still, the actual estimates are in practice the same.

2.1.2 MINQUE / MIVQUE

In 1971, Rao (Rao 1971*a*) (Rao 1971*b*) proposed an alternative estimating method for random effects models. This method now goes under the name Minimum Norm Quadratic Unbiased Estimation (MINQUE). This method minimizes an (Euclidean) norm under the restrictions that the estimator is a quadratic form of the observations and unbiased. In practice this method will need priors for both the random components and error term. Using the $N(x\beta, \Sigma)$ notation for the mixed effect model (found in section 1.1), we can write these priors in one covariance matrix $\Sigma^{(0)}$, which include both the random components and the error term.

For one linear combination of the random effects $a'\phi$, the estimation is done using the quadratic form $y'Qy$. For this quadratic form we let Q be

the symmetric matrix which minimizes $tr(Q\Sigma^{(0)}Q\Sigma^{(0)})$, under the conditions $QX = 0$ and $tr(Q\Sigma_t) = a_t$. Here Σ_t refers to the correlation matrix for the random effect corresponding to ϕ_t.

Swallow & Searle (1978), deduced several estimators based on this method for some common designs. As we see in their article, this method does not have do be solved by iteration. This makes the method more suited for larger datasets than the ML and REML estimators introduced in the next section. In addtion, the MINQUE does not require any normality assumptions, and is therefore useful even when we the underlying distributions are unknown.

In practice the most common priors are $\alpha = \mathbf{0}$ and $\varepsilon = 1$, as these priors leads to especially easy calculations. MINQUE with these priors are often referred to as MINQUE(0). Similarly, the term MINQUE(1), refer to MINQUE with priors $\alpha = \mathbf{1}$ and $\varepsilon = 1$.

If data are multinormally distributed, this estimator is the unbiased quadratic form estimator of the smallest variance. This has led to the term **mi**nimum **v**ariance **q**uadratic **u**nbiased **e**stimator (MIVQUE). In the balanced random effects setting, MIVQUE gives us the same estimates as the ANOVA estimates, regardless of the choice of priors. In the unbalanced case the equations are easily solved numerically.

MINQUE could also be solved recursively, letting the estimations of last iteration be the priors of the next iteration. Under normality assumptions this leads us to the REML estimate.

2.1.3 Maximum Likelihood (ML)

From section 1.1, we remember that the mixed effects model: $Y = X\beta + Z\alpha + \varepsilon$, could be expressed as a multivariate normal distribution $N(X\beta, \Sigma)$, where Σ includes both the variance of the random effects and error term. We now use the likelihood for the multivariate normal distribution, to find the likelihood for each individual in our mixed effects model:

$$L(\beta, \Sigma \mid \mathbf{X}) = |2\pi\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}\left[(x\beta-\mu)'\Sigma^{-1}(x\beta-\mu)\right]}$$

Because the individuals are independent, this means the joint likelihood becomes

$$L(\beta, \Sigma \mid \mathbf{X}) = \pi_{i=1}^{m} \left\{ |2\pi\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}\left[(x_i\beta-\mu)'\Sigma^{-1}(x_i\beta-\mu)\right]} \right\},$$

which gives the following log-likelihood:

$$L(\beta, \Sigma \mid \mathbf{X}) = -\tfrac{1}{2}\left\{m \log\left\{|2\pi\Sigma|\right\} + \sum_{i=1}^{m}\left[(x_i\beta - \mu)' \Sigma^{-1} (x_i\beta - \mu)\right]\right\}$$

If we maximize this for the balanced random effects model, we get the same estimates as we deduced for the ANOVA method. Still, for the unbalanced cases we get different estimates, and for these cases we do not get any explicit analytic formula. So for the unbalanced model we must use some sort of numerical methods for an interative solution of the equation. In practise this is often done by variants of the Newton-Raphson algorithm (Lindstrom & Bates 1988).

Even for quite complex models, modern computers usually find the Maximum Likelihood estimates pretty fast, as the likelihood in most cases is easy to maximize. One example of this is the likelihood for the Mississippi river's levels of nitrogen data mentioned in section 1.2. As we see from figure 2.1 and 2.2, it has a smooth one topped likelihood that is easily maximized.

2.1.4 Restricted Maximum Likelihood (REML)

As we shall see later on the Maximum Likelihood has one obvious weakness. Because of the estimation of the fixed effects, it gives a moderate bias for the random components. With this in mind, Patterson & Thompson (1971) suggested one modification of the Maximum Likelihood method to deal with the problem of biased estimates. The method has been named REML. This usually stands for Restricted Maximum Likelihood, but a more appropriate term would be Residual Maximum Likelihood, as it works on the residuals after estimation the fixed effects.

Instead of maximizing the likelihood as a whole, it divides the likelihood into two parts. This division is done so that the first part is invariant to the fixed effect $X\beta$, while the rest is dependent on $X\beta$. Using this, we can estimate the fixed effects from the part that is dependent on $X\beta$, and estimates for the random effects from the part that is invariant to $X\beta$. With this approach we can get unbiased estimates for the random effects, because we estimate the random effects without the problem of uncertain fixed effects. In practice this gives slightly adjusted estimates of the random components, while the fixed effect estimates remain intact.

From section 1.1, we remember that the mixed effects model can be written as $Y = N(X\beta, \Sigma)$, where Σ includes both the random effects and the

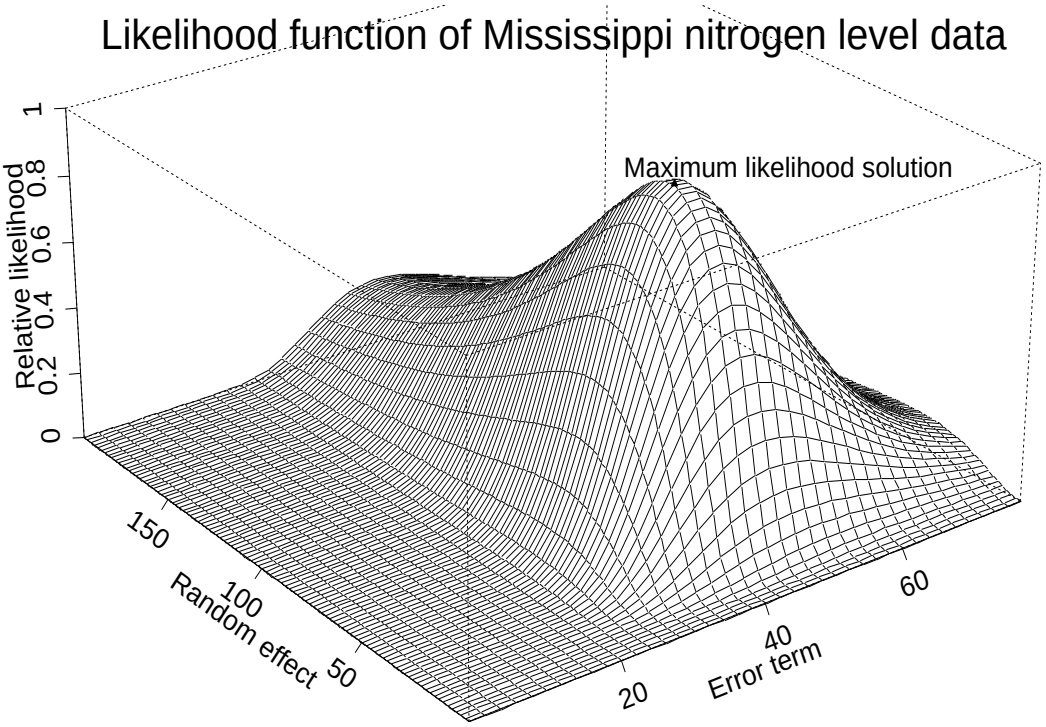


Figure 2.1: Likelihood funtion for the Mississippi river's levels of nitrogen data mentioned in section 1.2. Values are relative to the max value.

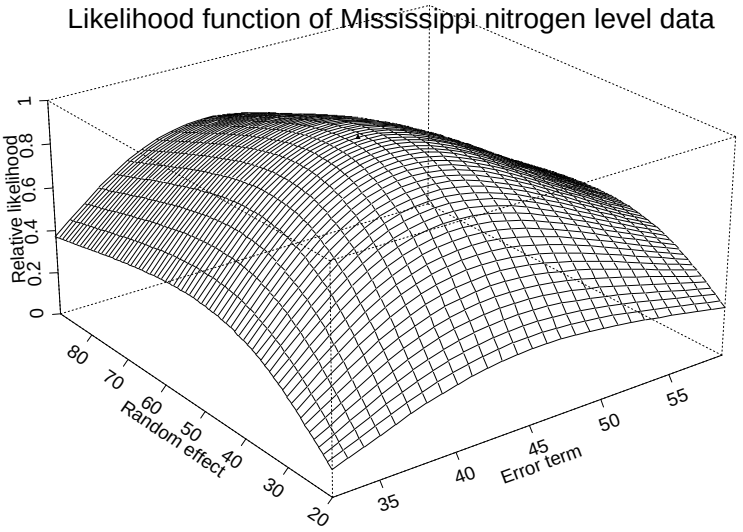


Figure 2.2: A closer view of the likelihood function seen in figure 2.1.

error term. One way of obtaining a likelihood for the random effects which is invariant to the fixed effect $X\beta$, is through the transformation $Y^* = AY$, where $A = I - X(X'X)^{-1}X'$. The new matrix Y^* is now singular of rank $N - p$, where N is the number of observations and p is the number of fixed parameters. Knowing this, we only would want to use $N - p$ rows of the matrix in the estimation. It can be shown (Searle & Henderson 1979) that the likelihood for this new data matrix Y^* is invariant to $X\beta$, and that the estimates remain the same for any choice of the $N - p$ rows from Y^*.

This transformation can also be presented with some other equations, like Hocking (1996), who deduced two new likelihoods. Still, both methods give the same estimates and the following log-likelihood for the random effects Σ:

$$L(\Sigma \mid \mathbf{X}) =$$
$$-\tfrac{1}{2}\log|\Sigma| - \tfrac{1}{2}\log|X'\Sigma^{-1}X| - \tfrac{n-p}{2}\log r'\Sigma^{-1}r - \tfrac{n-p}{2}\log\left[1 + \log\left(\tfrac{2\pi}{n-p}\right)\right]$$

Compared to the following likelihood for the ML approach:

$$L(\Sigma \mid \mathbf{X}) = -\tfrac{1}{2}\log|\Sigma| - \tfrac{n}{2}\log r'\Sigma^{-1}r - \tfrac{n}{2}\left[1 + \log(2\pi/n)\right],$$

where $r = y - X\left(X'\Sigma X\right)X^{-1}\Sigma^{-1}y$ and p is the rank of X. (See Littell et al. 1996.)

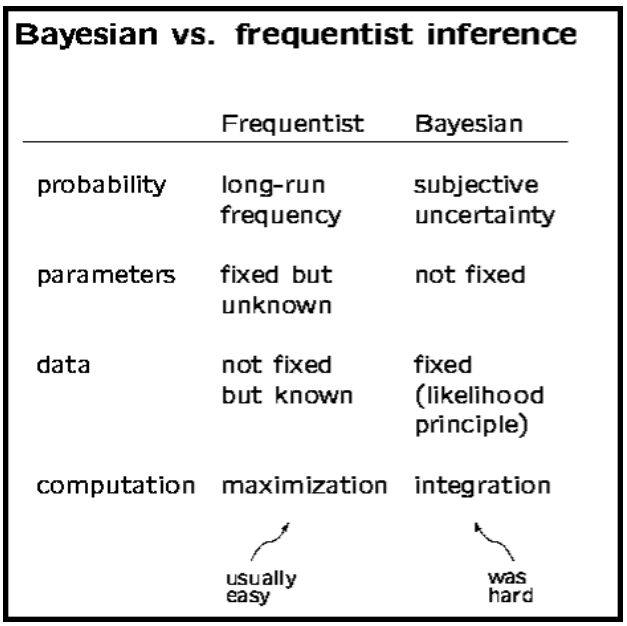


Figure 2.3: Difference of Bayesian and frequentist views. Slide from W. R. Gilks's BUGS course in Oslo, Norway, 1997.

As the Maximum Likelihood (ML) estimator, the Restricted Maximum Likelihood (REML) estimator is quite easy to find even for complex models using numerical methods combined with modern computers. For a discussion of an implementation of both the ML and REML estimates, see Harville1977.

2.1.5 Bayesian estimation: The Gibbs sampler

ANOVA, MINQUE, ML and REML are all called frequentist methods, as they assume that the estimates aim for true underlying parameters that have created the data through some sort of random process. One totally different approach is the Bayesian, where we assume that both observation (data) and model parameters are random quantities. With this we assume some sort of apriori distribution of the parameters, and use Bayes formula to find the estimates. This leads to different interpretations than the more traditional frequentist estimates we have seen earlier, something that is shown in figure 2.3.

With probability model $\rho(\theta|D)$ for data D, model parameters $\theta = (\theta_1, \theta_2, ..., \theta_K)$, and prior distribution of models parameters $\rho(\theta)$, we use Bayes's theorem:

$$\rho\left(\theta|D\right) = \frac{\rho(\theta)\rho(\theta|D)}{\int \rho(\theta)\rho(\theta|D)d\theta}$$

to find the posterior distribution $\rho(\theta \mid D)$. With this we can find estimates for any feature of the model, with:

$$E[f(\theta)|D] = \int f(\theta)\rho(\theta|D)\,d\theta$$

In many cases these features will be parameters, for which we can choose $f(\theta) = \theta$.

In practice, it is almost impossible to solve $E[f(\theta) \mid D]$ analytically for all but the most simple cases. This because we will have to solve very complicated and often multivariate integrations. This has long been the barrier for using Bayesian analysis in practice, but the recent advances in computer power have opened a new approach called Monte Carlo (MC) integration. Monte Carlo integration uses samples $\theta^1, \theta^2, ..., \theta^N$ from the posterior distribution $\rho(\theta \mid D)$, and a numerical estimate of $E[f(\theta) \mid D] = \sum_{i=1}^{N} f(\theta^i)$.

One question in Monte Carlo integration is how to draw the samples. The most obvious would be independent samples, but this is often complicated and very computer intensive. And it is really not compulsory to draw independent samples, as long as the Markov chain forgets its past after some repetitions. With this approach, we find the Markov chain via an transition kernel ρ and $\theta^{i+1} \sim \rho(\theta \mid \theta^i)$, where θ^{i+1} is independent of $\theta^1, \theta^2, ..., \theta^{i-1}$ given θ^i. If now all θ can be reached from θ^1 (irreducibility), the Markov Chain converges in distribution to its stationary distribution as $i \to \infty$.

Sampling directly from $\rho(\theta \mid \theta^i)$ can often be done with adaptive rejection sampling, but sometimes even that can be difficult and we may have to use the Metropolis-Hastings algorithm. In our study just plain Gibbs sampling as above works, so there is no need for the Metropolis-Hastings algorithm. A good source for learning more about Markov chain Monte Carlo methods is Gilks, Richardson and Spiegelhalter (1996).

2.2 Distributions of the estimates

We have now learned about several different estimation methods in variance and mixed effects models, but how well do these estimates perform in practice? And should some be favored for certain designs?

In the study of estimators performance there are three different approaches:

1. Theoretical studies of the estimates exact distributions.
2. Theoretical studies of the estimates limit distributions.
3. Simulations studies.

When available the first of these methods is usually preferred, as it gives precise results that are easily generalized for many parameter values. In the cases without any known exact distribution, we often have some limit distribution that are relevant to large sample data. This looks good on paper, but we have no general rule for what "large sample data" are, and many real world datasets are so small that the distribution of estimates often is very different from the theoretical large sample distribution.

These problems, and the fact that often even limit distributions are not available, lead us to the use of simulation studies. Also when limit distributions are available, simulation studies can be very important, as they can give us valuable information about the requirements for the limiting distribution. In addition, such studies can show us the distribution for small sample where the limit distributions do not work properly.

Historically, simulation studies have been hard to perform because of the calculations involved, but the revolution in computer power makes an increasing number of situations suited for simulations studies. Smart moves in the estimation and simulation process often increase the speed of this evolution.

2.2.1 Known theories for the distributions of the estimates:

Frequentist methods:

In the study of the performance of the frequentist estimating methods for mixed effects models, we will see use of all three methods for evaluating estimators discussed in the previous section. In this section we will present theory for exact and asymptotical distributions, while we later on will discuss result from simulations studies.

In the very basic setting of a balanced random effects model with normal distributions, we remember that REML, ANOVA and MINQUE (with all priors) all give the same estimates. In this case it is easy to compute the

exact distribution of the estimates, as we know the distributions of all the quadratic forms that are used in the estimation process. This is also true for ANOVA estimation in other balanced mixed effects models, as the quadratic forms in balanced mixed effects models have the following distribution.

The quadratic form $s_t = y'A_t y$, with expectation λ_t, has the following distribution in the balanced mixed effects model:

$\frac{s_t}{\lambda_t} \sim \chi^2_{(r_t,0)}$ for the random effects and error term,

where r_t is the degreas of freedom assiated with s_t quadratic form.

(e.g. Chi-square distribution with r_t degrees of freedom.)

$\frac{s_t}{\lambda_t} \sim \chi^2_{\left(r_t, \frac{1}{2\lambda_t}\alpha_t' X_t' \alpha_t X_t\right)}$ for the fixed effects,

where $\alpha_t X_t$ refers to the s_t specific part of the fixed effects $X\alpha$,
and r_t is the degreas of freedom assiated with s_t quadratic form.

(For proof see Hocking 1996.)

In the balanced random effects model, we remember from earlier examples that the ANOVA estimates are:

$$\hat{\sigma}_e^2 = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij}-\overline{y}_{i.})^2}{mn-m}$$

$$\hat{\sigma}_a^2 = \frac{\sum_{i=1}^{m}(\overline{y}_{i.}-\overline{y}_{..})^2}{(m-1)} - \frac{\hat{\sigma}_e^2}{n}$$

As we see, these estimates are based on the following two quadratic forms:

$$s_a = \sum_{i=1}^{m}(\overline{y}_{i.} - \overline{y}_{..}) \qquad \text{and} \qquad {}^2s_e = \sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij} - \overline{y}_{i.})^2,$$

where s_a is the between sum of squares, and s_e the within sum of squares.

Using the above theorem, this gives the following distributions for the quadratic forms:

$$\frac{s_a}{\lambda_a} = \frac{s_a}{\sigma_a^2 + \frac{\sigma_e^2}{n}} \sim \chi^2_{m-1} \qquad \text{and} \qquad \frac{s_e}{\lambda_e} = \frac{se}{\sigma_e^2} \sim \chi^2_{m*n-m},$$

where $\lambda_a = E(s_a)$ and $\lambda_e = E(s_e)$.

This leads to the following distributions for the estimates:

$$\hat{\sigma}_e^2 \sim \frac{\sigma_e^2}{mn-m}\chi^2_{m*n-m}$$

$$\hat{\sigma}_a^2 \sim \frac{\sigma_a^2+\frac{\sigma_e^2}{n}}{(m-1)}\chi^2_{m-1} - \frac{\sigma_e^2}{n(mn-m)}\chi^2_{m*n-m}$$

As we see, the distribution of $\hat{\sigma}_a^2$ is a combination of two chi-square distributions. In practice we have no tables for this distribution, but it can be evaluated using numeric integration.

As both the MINQUE (for any prior) and REML estimaties are equal to the ANOVA estimaties for this model, this distributions are also valid for the MINQUE and REML estimators in the one-way random effects model. Using the same theory, we can also find distributions for the estimates in other balanced designs. It can be shown that this is the minimum variance unbiased estimator, and therefore in some sense is an optimal estimator (For proof see Hocking 1996). As Maximum Likelihood here is just an adjustment of the denominator for the REML estimate, we could also easily deduce the distribution for the ML estimate.

This theory is also valid for balanced mixed effects and nested random effect models, but there are no known theory for the unbalanced cases. Still, we have some knowledge about some special cases:

- MINQUE is locally the unbiased quadratic estimator of minimum variance for the correct a priori values.
- The ANOVA method I and III give unbiased estimators, and for many models the variance of method I estimates are known (Searle 1971).
- For the Maximum Likelihood there is no known general exact distribution, but we could use the standard limit theory for ML-estimates. Under this theory the estimates are asymptomatically unbiased for large samples, and have the following distribution:

  $$\sqrt{n}(\widehat{\theta}_n - \theta) \rightarrow_d N\{0, J(\theta_0)^{-1}\}, \text{ where } J(\theta) = -E_\theta \frac{\partial^2 \log f(Y_i, \theta)}{\partial\theta\partial\theta'}.$$

  This is the general asymptomatical Maximum Likelihood theory for "large samples". In the simplest cases of fixed effect models, this just relates to a sample size n as $n \rightarrow \infty$, but in mixed effects models with

unbalanced designs it is not totally clear what this means, and we will need some specifications of the requirements for this large sample theory.

In Hartley & Rao (1967) such assumptions are described. The basic of these requirements is that the total number of observations shall increase to infinity in such a way that the number in each subclass shall always remain less than some "universal constant". For the unbalanced one-way random effects model seen in section 1.2, this will refer to the number of individuals increasing to infinity (e.g. $m \to \infty$). The goal of these assumptions is to ensure estimability, consistence and asymptotic efficiency.

As both ML and REML have the same limits as the sample sizes increases, this theory is also valid for the REML estimator. For further discussions about the minium requirements of the REML estimator in the mixed effects model, see Jiang (1995).

As $E_\theta \frac{\partial^2 \log f(Y_i,\theta)}{\partial\theta\partial\theta'}$ often are quite hard to find analytically, it is often evaluated numerically. This is especially practical when we solve the maximum likelihood equations with the Newton-Raphson algorithm, as the second derivatives of the log likelihood are available upon completion of the estimating process.

Still, it is sometimes possible to deduce $E_\theta \frac{\partial^2 \log f(Y_i,\theta)}{\partial\theta\partial\theta'}$ analytically, and Searle (1970) developed some very practical formulas for the linear mixed effects case.

2.2.2 Bayesian methods: The apriori distribution

For the Bayesian estimation methods, the a posteriori distributions are quite a different story than the distributions of frequentist estimators. Given a correct numerical solution of the equations, the a postriori distributions are the correct distributions for the given set of priors. If we have used known prior information for the priors, this gives us a correct picture. On the other side, in the situation when we have tried to supply uninformative priors, the a postriori distribution has no logical interpretation.

Still, we have some theory for the postrior distribution of non informative priors. If the priors are unifomly distributed over all the possible values, the

maximum of the postriori distribution will equal the ML estimates. This could easily be deduced for the Bayesian formula, by setting the prior distributions as uniform.

In all our earlier examples of random effect models, no such non informative priors are available to find. This as the parameter space for random effects usually span all positive values, and there is no uniform distribution that could span an infinite scale. Still, we can however in practise find priors that for all relevant values are pretty close to this aim. This again makes the distribution theory relevant as the number of observations increases, and the effects of the priors become less significant. How well this distribution work in small and moderate samples, will be dependent both on the number of observations and the choice of prior, and is something that should be investigated further in chapter 5.

2.3 Tests in mixed effects models

There are many different testing methods in mixed effects models , and recently there has even been written one book exclusively on this topic (Khuri, Mathew & Sinha 1998). In this section we will present the most commonly used methods, and look at some examples for the hypothesis of no variance component (e.g. H_0: $\sigma_a^2 = 0$) in the one-way balanced random effects model.

2.3.1 The Wald-test

The Wald-test is based on the asymptotically Maximum Likelihood theory from section 2.2.1. The test is in wide use in fixed effects models, and have proven to be quite useful for practical applications. It is included in most introductory courses in statistics, even thought it is often not given a specific name or accurate definition.

The Wald test takes the Maximum Likelihood estimate, and uses the Maximum Likelihood theory that tells us that this estimate is asymptotically normal distributed around the true value. Based on this, it then deduces a test statistic that can be evaluated by the normal distribution table. For the test that a parameters is zero, the test statistic will be on the form $\frac{estimate}{\sqrt{\widehat{var(estimate)}}}$ (E.g. estimate divided by its standard deviation).

As an example, we can use the very basic version of a fixed effects model: $x_i \sim N(\mu, \sigma)$ for $i = \{1, 2, 3, \ldots, n\}$. In this model the Wald test of H_0 :

$\mu = 0$, will have a test statistic of $\frac{\overline{x}}{\sqrt{n\widehat{\sigma}^2}}$, where $\overline{x} = \frac{1}{n}\sum_n^{i=1} x_i$ and $\widehat{\sigma}^2 = \frac{1}{n^2}\sum_n^{i=1}(x_i - \overline{x})^2$. Asymptotic Maximum Likelihood theory now tells us that this test statistic is standard normally distributed. In fact, exact distribution theory tells us that the test statistic is Student t distributed with n degrees of freedom.

For the hypothesis that a special variance components is not present, the test statistic will be the Maximum Likelihood estimate of the variance component divided by its estimated standard deviation. In many cases, this standard deviation is unknown, but an approximation can always be found numerically by the asymptotic variance of the estimate (see section 2.2.1). In practice this approximation has become the standard method, as it is easy to implement in statistical software packages.

As an example we can look at the balanced one-way random effects model (see section 1.2). In this model we can get variance estimates for the estimates using both the "Classical" asymptotically based method, and exact distribution theory. Both these approaches are shown below:

Exact formula for the $\hat{\sigma}_a^2$ variance (under ML-estimation):

From section 2.1.1, we remember that the Maximum likelihodd estimated in a one-way balanced random effects model is $\hat{\sigma}_a^2 = \frac{\sum_{i=1}^m (\overline{y}_{i.} - \overline{y}_{..})^2}{m} - \frac{\hat{\sigma}_e^2}{n}$. Using this we find that

$$\begin{aligned}\text{var}\left(\hat{\sigma}_a^2\right) &= \text{var}\left(\frac{\sum_{i=1}^m (\overline{y}_{i.} - \overline{y}_{..})^2}{m} - \frac{\hat{\sigma}_e^2}{n}\right) \\ &= \text{var}\left(\frac{n^2 s_1}{m} - \frac{s_e}{n*(mn-m)}\right),\end{aligned}$$

where $s_{1=}\sum_{i=1}^m (\overline{y}_{i.} - \overline{y}_{..})^2$ and $s_e = \sum_{i=1}^m \sum_{j=1}^n (y_{ij} - \overline{y}_{i.})^2$.

Because of independence of s_1 and s_e (see Hocking 1996), we get that

$$\begin{aligned}\text{var}\left(\frac{s_1}{m} - \frac{s_e}{n*(mn-m)}\right) &= \text{var}\left(\frac{s_1}{m}\right) + \text{var}\left(\frac{s_e}{n*(mn-m)}\right) \\ &= \frac{1}{m^2}\text{var}\left(s_1\right) + \frac{1}{n^2*(mn-m)^2}\text{var}\left(s_e\right).\end{aligned}$$

To find the variance of s_e, we use the distribution theory found in section 2.2.1, which gives that

$\frac{s_e}{\sigma_e^2} \sim \chi^2_{mn-m}$.

This means that $\text{var}\left(\frac{s_e}{\sigma_e^2}\right) = 2(mn-m)$, and solving this with regard to $\text{var}(s_e)$, we get that $\text{var}(s_e) = 2(mn-m)\sigma_e^4$.

As for $\text{var}(s_e)$, we again use distribution theory from section 2.2.1, and find that

$\frac{s_1}{\sigma_a^2+\frac{\sigma_e^2}{n}} \sim \chi^2_{m-1}$.

This gives $\text{var}\left(\frac{s_1}{\sigma_a^2+\frac{\sigma_e^2}{n}}\right) = 2(m-1)$, and solving with regard to $var(s_1)$ we get

$$\begin{aligned}\text{var}(s_1) &= 2(m-1)\left(\sigma_a^2+\frac{\sigma_e^2}{n}\right)^2 \\ &= 2(m-1)\left(\sigma_a^4+2\frac{\sigma_e^2\sigma_a^2}{n}+\frac{\sigma_e^4}{n^2}\right).\end{aligned}$$

Now we can use these results to find $var(\hat{\sigma}_a^2)$:

$$\begin{aligned}\text{var}\left(\hat{\sigma}_a^2\right) &= \frac{1}{m^2}v\text{var}(s_1)+\frac{1}{n^2*(mn-m)^2}\text{var}(s_e) \\ &= \frac{1}{m^2}2(m-1)\left(\sigma_a^4+2\frac{\sigma_e^2\sigma_a^2}{n}+\frac{\sigma_e^4}{n^2}\right)+\frac{1}{n^2*(mn-m)^2}2(mn-m)\sigma_e^4 \\ &= 2\frac{m-1}{m^2}\left(\sigma_a^4+2\frac{\sigma_e^2\sigma_a^2}{n}+\frac{\sigma_e^4}{n^2}\right)+\frac{2\sigma_e^4}{n^2*(mn-m)}.\end{aligned}$$

Asymptotic variance for $\hat{\sigma}_a^2$ (under ML-estimation):

From section 2.2.1, we remember that we could get an approximation for the variance of $\hat{\sigma}_a^2$ by using general asymptotic theory for the maximum likelihood estimates:

$$\sqrt{n}(\widehat{\theta}_n-\theta) \rightarrow_d N\{0, J(\theta_0)^{-1}\} \text{ where } J(\theta) = -E_\theta\frac{\partial^2 \log f(Y_i,\theta)}{\partial\theta\partial\theta'}$$

For this model asymptotical theory here refers to the number of individuals increasing to infinity (e.g. $m \rightarrow \infty$). For further information about asymptotical assumptions in mixed effects models, see Hartley & Rao (1967).

To find $-E_\theta \frac{\partial^2 \log f(Y_i,\theta)}{\partial\theta\partial\theta'}$, we used that the likelihood for the one-way balanced random effects model. Written on the alternative $N(x\beta, \Sigma)$ form described in section 1.1, this likelihood become:

$$f(x) = |2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\},$$

where $\Sigma = \sigma_a^2 D_m(J'J) + I_{n*m}\sigma_e^2$ and J is a vector of n ones.

This gives a log-likelihood of

$$\log f(x) = {}^{-\frac{1}{2}} \log|2\pi\Sigma| + \left\{-\tfrac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}.$$

Taking the derivative twice, we get that

$$\begin{aligned}\frac{\partial^2 \log f(Y_i,\theta)}{\partial\theta\partial\theta'} &= \frac{\partial^2\left\{-\frac{1}{2}\log|2\pi\Sigma|\right\}}{\partial\theta\partial\theta'} + \frac{\partial^2\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}}{\partial\theta\partial\theta'} \\ &= \frac{\partial^2\left\{-\frac{1}{2}\log|2\pi\Sigma|\right\}}{\partial\theta\partial\theta'} - \frac{1}{2}tr\left[\Sigma^{-1}\frac{\Sigma\partial}{\partial\sigma_a^2}\Sigma^{-1}\frac{\Sigma\partial}{\partial\sigma_a^2}\right].\end{aligned}$$

For finding $\Sigma^{-1}\frac{\Sigma\partial}{\partial\sigma_a^2}\Sigma^{-1}\frac{\Sigma\partial}{\partial\sigma_a^2}$, we use that

$$\tfrac{\Sigma\partial}{\partial\sigma_a^2} = D_m(J'J) \text{ and } \Sigma^{-1} = I_{n*m}\alpha + \beta(D_m(J'J) - I_{n*m}),$$

where $\alpha = \frac{(n-1)\sigma_e^2+\sigma_a^2}{\sigma_e^2(n\sigma_e^2+\sigma_a^2)}$ and $\beta = \frac{\sigma_a^2}{\sigma_a^2(n\sigma_e^2+\sigma_a^2)}$.

This gives us

$$\Sigma^{-1}\tfrac{\Sigma\partial}{\partial\sigma_a^2} = \tfrac{1}{n\sigma_e^2+\sigma_a^2},$$

and

$$\Sigma^{-1}\tfrac{\Sigma\partial}{\partial\sigma_a^2}\Sigma^{-1}\tfrac{\Sigma\partial}{\partial\sigma_a^2} = D_m(J'J)\tfrac{n}{(n\sigma_e^2+\sigma_a^2)^2}.$$

This means that

$$tr\left[\Sigma^{-1}\tfrac{\Sigma\partial}{\partial\sigma_a^2}\Sigma^{-1}\tfrac{\Sigma\partial}{\partial\sigma_a^2}\right] = m*n\left[\tfrac{n}{(n\sigma_e^2+\sigma_a^2)^2}\right] = \tfrac{nm^2}{(n\sigma_e^2+\sigma_a^2)^2}.$$

And finally we now have that the asymptotic variance for $\hat{\sigma}_a^2$ is

$$\begin{aligned} var\left(\hat{\sigma}_a^2\right) = J(\theta)^{-1} &= \left[-E_\theta \frac{\partial^2 \log f\left(Y_i,\theta\right)}{\partial\theta\partial\theta'}\right]^{-1} \\ &= \left[-\left(-\frac{1}{2}\right) tr\left[\Sigma^{-1}\frac{\Sigma\partial}{\partial\sigma_a^2}\Sigma^{-1}\frac{\Sigma\partial}{\partial\sigma_a^2}\right]\right]^{-1} \\ &= \frac{2}{tr\left[\Sigma^{-1}\frac{\Sigma\partial}{\partial\sigma_a^2}\Sigma^{-1}\frac{\Sigma\partial}{\partial\sigma_a^2}\right]} = \frac{2}{\left(\frac{nm^2}{(n\sigma_e^2+\sigma_a^2)^2}\right)} \\ &= \frac{2\left(\sigma_a^4 + 2\frac{\sigma_e^2\sigma_a^2}{n} + \frac{\sigma_e^4}{n^2}\right)}{m}. \end{aligned}$$

To summarize the two methods for calculation of $var(\hat{\sigma}_a^2)$, we have an exact variance given by

$$var\left(\hat{\sigma}_a^2\right) = 2\tfrac{m-1}{m^2}\left(\sigma_a^4 + 2\tfrac{\sigma_e^2\sigma_a^2}{n} + \tfrac{\sigma_e^4}{n^2}\right) + \tfrac{2\sigma_e^4}{n^2*(mn-m)},$$

while the asymptotic variance is

$$var\left(\hat{\sigma}_a^2\right) = \tfrac{2}{m}\left(\sigma_a^4 + 2\tfrac{\sigma_e^2\sigma_a^2}{n} + \tfrac{\sigma_e^4}{n^2}\right).$$

Remark: The same techniques could be used for the Wald test build on the REML estimator, causing only in small corrections of the final results.

This again gives us the following test statistics:

$$\text{Exact:} \quad \frac{\hat{\sigma}_a^2}{\sqrt{2\tfrac{m-1}{m^2}\left(\sigma_a^4 + 2\tfrac{\hat{\sigma}_e^2\hat{\sigma}_a^2}{n} + \tfrac{\hat{\sigma}_e^4}{n^2}\right) + \tfrac{2\hat{\sigma}_e^4}{n^2*(mn-m)}}}$$

$$\text{Asymptotic:} \quad \frac{\hat{\sigma}_a^2}{\frac{2}{m}\left(\hat{\sigma}_a^4 + 2\frac{\hat{\sigma}_e^2\hat{\sigma}_a^2}{n} + \frac{\hat{\sigma}_e^4}{n^2}\right)}$$

Comparing these test statistics, we see that as m increases $2\frac{m-1}{m^2} \to \frac{2}{m}$. In addition will the $\frac{2\hat{\sigma}_e^4}{n^2*(mn-m)}$ term usually be quite small, so the asymptotic variance will in most cases make a fairly good estimate of the true variance.

In real world examples the asymptotical approximation is also especially easy and effective to implement, as we in Maximum Likelihood estimation already get $\frac{\partial^2 \log f(Y_i,\theta)}{\partial\theta\partial\theta'}$, when maximizing with the Newton-Raphon algorithm. In practice it is also implemented in many statistical packages like the SAS system.

For the standard definition of the random effects model seen in section 1.2, the null hypothesis $\sigma_a^2 = 0$ should refer to a boundary point of the parameter space. This should at first look indicate that the asymptotic theory will have boundary problems. Still, we can show that this is not the case, as an equivalent hypothesis do not have these problems.

Remember the alternative definition of the mixed effects model in section 1.1. Here we could write the model as $N(x\beta, \Sigma)$, where Σ is the variance-covariance matrix of both the random effects and the error term. For the one-way balanced random effects model, this gives a covariance matrix of $\Sigma = \sigma_e^2 I_{n*m} + \sigma_a^2 D_m\,(J'J)$, where J is a vector of n ones. Writen on this form, σ_a^2 can be defined to take both positive and negative values. Using this defintion, the test $\sigma_a^2 \leq 0$ should not result in any boundary problems. For positive values of the estimates, the test statistic for this model will be identical to the one in the standard definition of the model. This again means that it also take the same asymptotic distributions for positive values. Since the asymptotical theory is valid for this alternative model, the theory will also be valid in the $\sigma_a^2 = 0$ test for the standard definition of the model, all rejections of the test is positive estimates.

2.3.2 Likelihood ratio based testing

In likelihood based testing, we use the following asymptotic large sample distribution:

Let l_0^* be the log maximum likelihood under the H_0 restriction, and l_1^* be the maximum likelihood solution without this restriction. Then

$$-2\log\lambda = 2(l_1^* - l_0^*) \rightarrow_d \chi_p^2 \text{ under } H_0,$$

where $\lambda = l_0^*/l_1^*$, and p is the differences in dimensions of the parameter space with and without the H_0 restriction.

(E.g. Under H_0, $-2\log\lambda$ converges to a chi-square distribution with p degrees of freedom when the number of observations increases.)

Asymptomatical theory here have the same assumptions as in section 2.2.1. For the unbalanced one-way random effects model (see section 1.2), this will refer to the number of individuals increasing to infinity (E.g. m$\rightarrow \infty$).

Using numerical calculations, it is generally fairly straightforward to find l_0^* and l_1^*. This is especially true when we use Maximum Likelihood estimates, as we then already have found values for the unrestricted maximum likelihood. Still, this method uses quite a bit computer power, as we need to do perform several maximum likelihood estimations.

For the one-way random effects model, the hypothesis of no random component refers to H_0: $\sigma_a^2 = 0$. This gives $p = 1$, l_1^* as the log-likelihood under the one-way random effects model and l_0^* is the log-likelihood for a fixed effects model with just one constant term and a normal distributed error term.

2.3.3 The F-test

In the cases where we know the distributions of the quadratic forms, it is also possible to deduce a test statistics with known distribution under H_0. This includes most balanced designs, using distribution theory from section 2.2.1. Using the known distributions under H_0, we then can deduce tests with exact p-values under H_0. Since the test statistic follows the F distribution, these tests are called F-tests.

As an example, we can take the balanced one-way random effects model from section 1.2. Using the within and between sums of squares from section from 2.1.1, and the distribution theory from section 2.2.1, we now get the exact distributions of these sums of square:

$$\frac{s_a}{\lambda a} = \frac{s_a}{\sigma_a^2 + \frac{\sigma_e^2}{n}} \sim \chi^2_{m-1}$$

$$\frac{s_e}{\lambda e} = \frac{se}{\sigma_e^2} \sim \chi^2_{m*n-m}$$

Combining these, we get the test statistic

$$\frac{s_a/(m-1)}{s_e/(m*n-n)},$$

which under $H_0 : \sigma_a^2 = 0$ take the a $\frac{\lambda_a}{\lambda_e} F\left(m-1, m*n-n\right)$ F-distribution. This gives us a test with a correct significant level under H_0. In some sense this test is optimal, but we do still not know its power under Ha.

As we see, these tests are easily deduced in most balanced designs, but in most unbalanced design such tests impossible to deduce using exact distribution theory. In some of these cases it is possible to deduce approximate tests, with some modifications of the corresponding test in the balanced design (see Hocking 1996). One example of this is the unbalanced one-way random effects design, where we can build an approximate test from the corresponding test in the balanced design.

Chapter 3

Estimation methods in practical use

> The mathematician, carried along on his flood of symbols,
> dealing apparently with purely formal truths,
> may still reach results of endless importance for
> our description of the physical universe.
> Karl Pearson (1857-1936)

After learning about the different models, two questions arise:

- Which models are in practical use today?
- Does it really matter what kind of model we use?

3.1 Estimation methods in common statistical packages

In most practical cases, all the estimation methods mentioned in the previous chapter require quite a lot of calculations. Historically, this has restricted the use of the mixed effects models among researchers, but in the last decade mixed effects models have been included in many of the most popular statistical packages and the use has flourished. Today, probably almost all applications of the model is done with one of the preprogrammed packages, and these packages should give us a good idea of which estimation methods are in daily use.

Statistical package	ANOVA	MINQUE	ML	REML
R (0.64)	-	-	**Lme** (Beta version)	Lme (Beta version)
S-PLUS (4.5 / 5)	-	**Varcomp** (0,1)	**Lme** & varcomp	Lme & varcomp
SAS (6.12)	Proc Varcomp (Type I)	Proc **Varcomp** (0,1)	Proc Mixed & Varcomp	Proc **Mixed** & Varcomp
SPSS (9.0)	Type I & III	**Available** **(0,1)** & (1,1)	Available	Available
Statistica (99)	**Type I, II & III**	Available	Available	Available

Table 3.1: Mixed effects estimation methods in some common statistical packages - March 1999. Default methods in bold font, and command line routines are given by its name (like Lme, Varcomp, etc.). As we see, both SAS and S-PLUS have two different set of routines for mixed effects models. Available priors for MINQUE are shown in parenthesis.

Frequentist methods

Mixed effects models have traditionally been analyzed using different sorts of frequentist methods. In table 3.1, we see which frequentist methods that are implemented in some of the most common statistical programs. Readers should be especially aware of the default methods, as many users probably apply these methods without further notice. As we can see in table 3.1, the default methods vary between all the different models. This probably means that all methods are in widespread practical use.

For practical use today, SPSS is one of the most user friendly packages with a well organized graphical user interface. Still, it has its limitations, with a somewhat limited output and a restriction to just random effect models. For mixed effect models and a more comprehensive output, we have to go with other packages. Among these, SAS has a very comprehensive output,

Statistica provides easy analyzing through a graphical user interface, and S-PLUS delivers especially good graphics (such as trellis graphs) and a very flexible and powerfull set of modeling tools.

In the fast developing field of mixed models, there are also many new software packages on its way, that for the time being are in beta releases. Among the most interesting are the new NLME 3.0 library for S-PLUS (Pinheiro & Bates 1997), which will be integrated in the upcoming S-PLUS 2000 release (due summer 1999). These new routines promise easier use, new graphics, comprehensive output, and methods for analyzing an even wider sets of models.

Other interesting upcoming releases includes a port of the S-PLUS NLME library (including lme-routines) to the open source R system. This will give users of both Linux, UNIX and Window a powerfull free package for analyzing mixed effect models.

All in all, mixed effect models is for the time being one of the fastest developing fields in statistical computing. For further information on these packages, see references on page 115. In addition to these packages, we also have many other packages with mixed effects models such as Minitab, BMDP and MLwiN.

Bayesian methods

For a long time, Bayesian estimation has mostly been of theoretical interest, but recent progress in computing power and software has opened for practical use of these alternative estimation methods. One of the first such revolutions came with the BUGS package. This is a freeware program for Bayesian analyzing using the Gibbs sampler, which Cambrigde MCR Biostatistics Unit released for several operation systems during the first half of the nineties. This package opened for easy Bayesian analysis through a command line interface, and with the commercial release of Winbugs version 1.0 in 1998, we also got a Bayesian package with a graphical interface.

With these new Bayesian software packages, Bayesian methods have already come in pretty widespread use, and in the future Bayesian methods will probably also be included in several standard statistical packages. For the mixed effects models, these methods often give especially easy modeling properties (Gilks et al. 1993). This has given random and mixed effect models a special position as one of the premier examples of practical applications of Bayesian methods like the Gibbs sampler.

For further information about the BUGS package, see reference on page 115.

3.2 Different estimation methods - One numerical example

As we now have seen, the statistical packages implement different estimation methods, but does this have any practical consequences?

To shed some light on this, we have used six different estimation methods on one dataset. Two of these methods, the MINQUE and Gibbs sampler, require priors. For the MINQUE, we have chosen the two "classical" priors. Both these priors have $\sigma_e^2 = 1$, but one with $\sigma_a^2 = 0$ and the other with $\sigma_a^2 = 1$. These are called MINQUE(0,1) and MINQUE(1,1).

When it comes to the Gibbs sampler, we have also chosen two sets of priors. The first of this, Gibbs sampling - "Default", is taken from the random effects example in the BUGS creators examples collection (Spiegelhalter, Thomas, G. & Gilks 1996). These are $\mu \sim N(0, 1e{+}10)$, $\sigma_e^2 \sim \Gamma(0.001, 0.001)$ (Here Γ refers to the gamma distribution) and $\sigma_a^2 \sim \Gamma(0.001, 0.001)$. Remember that $\Gamma(r, \lambda)$ has an expectation of $\frac{r}{\lambda}$, and variance of $\frac{r}{\lambda^2}$, so that this means the prior for both the variance component and error term has expectation 1 and variance 1000. As an alternative we have also used the prior $\mu \sim N(0, 1e + 10)$, $\sigma_e^2 \sim \Gamma(1, 1)$ and $\sigma_a^2 \sim \Gamma(1, 1)$. This gives the priors for both the variance component and error term an expectation and variance of one. We will discuss this choice of priors further in section 5.2.2. Using these priors, each Gibbs sampler estimate is the mean of 10 000 steps from the Monte Carlo chain.

These methods are then used on the nitrogen levels in the Mississippi river data shown in section 1.2. This resulted in the estimates in table 3.2, and as we can see the estimates vary considerably. This motivates studies of the estimates performance, so we can give users some practical advice on which methods to use on their data. In the next two chapters, we shall have a look at some earlier studies and perform a new one.

Estimation method	Estimated variance of random effect (σ_a^2)	Estimated varince of error term (σ_e^2)
ML	51.3	42.7
REML	63.3	42.7
MINQUE(0,1)	45.8	51.4
MINQUE(1,1)	62.6	42.7
ANOVA	56.2	42.6
Gibbs sampling - "Default"	103.1	46.2
Gibbs sampling - Alternative	57.2	45.3

Table 3.2: Estimates for the Mississippi nitrogen level data. See the text for information about which priors are used for the Gibbs sampling estimates.

Chapter 4

An overview of earlier Monte Carlo studies

> The number of transistors on a chip doubles every two years.
> Dr. Gordon Moore (1929 -)

In the last chapter, we saw an example of how much the different estimates can vary in mixed effects models. This raises several questions about the estimators performance, and which estimators are recommended for practical use in different situations. As we saw in section 2.2, there are three different approaches for studying estimators performance: Theoretical studies of the exact distributions, use of asymptotical (limit) distributions and simulation studies. We also learned that there were no known exact distribution for many of the estimators in some of the most central mixed effect model designs. For some of these cases, we had limited distributions, but even then, we often do not know how these apply for small samples. This leaves us with the third solution, simulation studies, and in this chapter we shall have a look at some earlier simulation studies of methods for analyzing mixed effect models.

Early days - Numerical studies

For a long time, simulation studies in mixed effects models were practically impossible, because of the large calculations involved. This can easily be understood by noting the fact that even on moderate data sets, Maximum Likelihood (ML) estimations were often not advisable in practice, since the calculations grew too large for the available computing power.

In the seventies and eighties, this began to change as the computers grew more powerful and became widely available. This soon resulted in some sim-

ulation studies such as Maddala & Mount (1973) and Miller (1979). Maddala & Mount (1973) looked at methods of estimating the slope coefficient in the mixed effects models, while Miller (1979) compared ML and REML estimators in the two-way balanced random effects model.

We also saw some articles deducing formulas for expectations and variances in some special cases, and then using this to evaluate some estimators. The first such study was done by Swallow and Searle in 1978, in which they compared ANOVA and MINQUE estimators for a one-way random effects model with unbalanced data. From earlier, we remember that both these estimators are unbiased, so they could use the variances as a reasonable measurement of the estimators performance. Swallow and Searle deduced formulas for these variances, and applied them to 13 different unbalanced one-way random effect model designs, with $\sigma_a^2 = \{1/2, 1, 5, 10, 20\}$ and $\sigma_e^2 = 1$.

In this study they found that:

- In many situations, the variation of the ANOVA variance component estimates (of σ_a^2) is far from the lower bound of the MINQUE estimator. In practice, the MINQUE with prior $\sigma_a^2 = \sigma_e^2 = 1$, is significantly better in estimating the variance component (σ_a^2) than the ANOVA estimates. A notable exception is for a small variance component, where there is little difference at $\sigma_a^2 = 1/2$.

- For the ANOVA estimates of the error term (σ_e^2), the variance is very near to the lower bound for an unbiased estimator. With $\sigma_a^2 = \sigma_e^2 = 1$ as priors, the MINQUE estimates perform significantly worse for large variance components (σ_a^2), but for small and moderate values the MINQUE with these priors gives results very near the theoretical lower bound for an unbiased estimator.

Three years later, Swallow followed up with a new article (Swallow 1981). In this article, he looked at the same design as in 1978, but with several different alternatives of priors for the MINQUE estimator. He then found that when the true σ_a^2/σ_e^2 is not too small (e.g. larger than ≈ 1), and the ratio of the prior variance component versus prior error term (e.g. $\sigma_{a0}^2/\sigma_{e0}^2$) is not severely underestimated, the MINQUE estimate has a variance very near the lower bound and is more efficient than the ANOVA estimator.

This leads to the following advice for applied statistics:

In the choice between ANOVA and MINQUE estimates in unbalanced one-way classification designs, MINQUE is usually a good choice when $\sigma_a^2/\sigma_e^2 > 1$, but we should be careful not to underestimate the ratio of the prior variance component versus prior error ($\sigma_{a0}^2/\sigma_{e0}^2$). This is especially true if we are interested in estimation of the variance component (σ_a^2), while the estimation of the error term (σ_e^2) is somewhat more robust. It can be shown (Hartley, Rao & LaMotte 1978) that MINQUE with $\sigma_{a0}^2 = 0$ gives especially easy computations. This has led to the MINQUE(0)[1] being implemented in many statistical packages, but Swallow's (1981) study shows that this is a dangerous estimator for data with large σ_a^2/σ_e^2.

4.1 The first large simulation study: Swallow & Monahan

The first comprehensive simulation study was done as early as 1984 by Swallow and Monahan. In this article they compared ANOVA, MINQUE(0), MINQUE(A)[2], REML, ML, and the adjusted ML[3] estimator.

The study was made possible through relatively fast computers and a smart trick. Since the subgroup means and subgroup sum of squares are sufficient for the estimators, it is not necessary to simulate individual data. Instead, they simulated the subgroup means and subgroup sum of squares, and used these values to find the estimates. In addition, they dropped every set of data where the REML and ML routines failed to converge in 20 interactions. This happend very rarely, but still means quite a lot for the total computating time. All in all, this article is very interesting and illuminating, and became **the** breakthrough article about the efficiency of different estimators in random effects models.

As we in most practical applications will suppose that all variance components are positive (e.g. $\sigma_a^2 > 0$), Swallow & Monahan have chosen to set all negative values to zero.

For the estimation of the variance component (σ_a^2), Swallow & Monahan find the following:

[1] MINQUE(0) is the MINQUE estimator with priors $\sigma_{a0}^2 = 0$ and $\sigma_{e0}^2 = 1$.

[2] MINQUE(A) is the MINQUE estimator with the ANOVA estimates as prior.

[3] ML with adjustment of degrees of freedom, in an attempt to solve the problems of bias.

- In most cases, all the estimators have only a moderate bias. One exception is the ML estimator, which has a significant downward bias for large values of σ_a^2. In addition, all the other estimators take a substantial upward bias in most cases with small variance components (σ_a^2). This is contributed by the truncation of negative values, which ruins the theory that both REML, MINQUE(0), MINQUE(1) and ANOVA are unbiased.

- For large variance components (e.g. large σ_a^2/σ_e^2), ANOVA and especially the MINQUE(0) estimator, have a considerable larger variance than the other estimates.

- The ML estimator often has a significantly lower variance than all the other estimates. This can partly be attributed to the combination of underestimation and truncation of negative values. The bias adjusted ML-estimate is less influenced by this, and in most cases takes a substantially larger variance than the standard ML estimate.

- For some severely unbalanced designs, the REML and to some degree the ANOVA, ML and MINQUE(A) estimates have a considerably larger variance than the theoretically lower bound for an unbiased estimator.

- In most cases not mentioned earlier, all the estimators have variances quite near the lower bound for an unbiased estimator. For large variance components (σ_a^2), the variance is generally slightly over the lower bound. On the other hand, estimators for small variance components (e.g. < 1) sometimes (even) dip below the lower bound. This especially low variance can be attributed to the truncation of negative values, which we remember also gives us somewhat biased estimates even with the REML, MINQUE(0), MINQUE(A) and ANOVA methods.

For the estimation of the error term (σ_e^2), Swallow & Monahan found the following:

- All estimators have very little bias.

- All estimates have variances near the lower bound for an unbiased estimator, except MINQUE(0), which becomes a very poor estimator as the variance component and number of individuals (N) increases. Comparing the estimators, they also found a tendency for a somewhat lower variance of the ML estimates.

As for the speed of convergence, Swallow & Monahan found that both ML and REML usually converge very rapidly. With todays relatively fast computers, this is no longer so important, but for especially large datasets the MINQUE(0) can be a reasonable solution.

4.2 Other studies

After Swallow & Monahan (1984), several simulation studies have been published. Some of these have been simulation studies of special cases or non-standard estimators, while most studies have been a combination of an application and some related simulations. An example of the first is Yu, Searle & McCulloch 1994, who among other things look at the maximum likelihood estimator for non-normal distributions.

An example of a study combining an application with some simulation results is Giesbrecht & Burns (1985). In this article, they developed a two-stage analysis based on a mixed effects model, and they have done a moderate amount of simulations on this new theory. Another such study is Engel & Buist (1996), who look at an alternative to maximum likelihood estimation in a mixed effects model.

4.3 Still unsolved questions

Although there have been quite a lot of simulation studies in recent years, there are still several central questions which remain unanswered. These include:

1. In addition to the variance and mean, how are the other properties of the estimators, like the distributions that are used for confidence intervals and testing?

2. How well does the Gibbs sampler with "uninformative priors" perform in comparison to the other estimates?

3. Which testing methods are best for mixed effects models?

In the next two chapters, we will try to shed some light on these questions.

Chapter 5

A Monte Carlo study of the different estimation methods

> The pure and simple truth is rarely pure and never simple.
> Oscar Wilde (1854-1900)

As we have seen in the previous chapter, there are still several unanswered questions about the properties of the different estimating methods used for mixed effects models. In this chapter, we will present a new study, which aims at some of these questions. As a start, we will repeat most of Swallow & Monaham (1984), but take a somewhat more in-depth look at the estimators properties. Later, we also will include the increasingly popular Gibbs sampler estimating method.

Designs used in this study

In this study, we have limited the models to different designs of the one-way random effect model. This model was chosen as it is both in wide use, and it is the simplest and most perspicuous mixed effects model. Still, it can also properly give us some hints of what happens in the more complex models like multilevel models and mixed effects with a random regression coefficient. This must, however, be examined further, and is a good theme for new studies. As we remember from section 1.2, the model can be written as:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

where:

- μ is a fixed constant.

Design number	Number of observations
1	$m = 3$, $n_i = \{3, 5, 7\}$
4	$m = 6$, $n_i = \{3, 3, 5, 5, 7, 7\}$
11	$m = 9$, $n_i = \{1, 1, 1, 1, 1, 1, 1, 19, 19\}$
12	$m = 3$, $n_i = \{2, 10, 18\}$
20	$m = 3$, $n_i = \{3, 6, 6\}$
21	$m = 6$, $n_i = \{3, 3, 6, 6, 6, 6\}$

Table 5.1: Unbalanced designs included in this study.

Design number	Number of observations
1	$m = 3$, $n_i = \{2, 2, 2\}$
3	$m = 3$, $n_i = \{5, 5, 5\}$
6	$m = 6$, $n_i = \{5, 5, 5, 5, 5, 5\}$
7	$m = 10$, $n_i = \{2, 2, 2, 2, 2, 2, 2, 2, 2, 2\}$
9	$m = 10$, $n_i = \{5, 5, 5, 5, 5, 5, 5, 5, 5, 5\}$

Table 5.2: Balanced designs included in this study.

- α is the random effect with $\alpha_i \sim N(0, \sigma_a^2)$.
- ε_{ij} is the variation of the observation (as sampling error etc.), with $\varepsilon_{ij} \sim N(0, \sigma_e^2)$.
- $i = 1, ..., m$ and $j = 1, ..., n_i$

(An example of this model in practical use is shown is section 1.2.)

This is the general model, under which we have simulated with $\mu = 3$ and $\sigma_e^2 = 1$ for different values of m, σ_a^2 and n_i. We have chosen fix μ at a specific value, but this choice has no influence on the random component estimators. As for σ_e^2, it has some effect, but only through the ratio of σ_a^2/σ_e^2 and for the general scale of the estimates.

Varying m and n_i, we have constructed several different small and medium sized designs. These designs aim to span a wide range of situations commonly found in practical work, while still also giving us some ideas about how different values of m and n_i affect the different estimates. To ease this work, we have given each balanced and unbalanced design a unique number. The designs mentioned in this study are shown in table 5.2 and 5.1, where the number of observations refer to $n_1, n_2, ..., n_m$. For the unbalanced design number 1, this means that $n_1 = 3$, $n_2 = 5$ and $n_3 = 7$, while $m = 3$. E.g. three observations, with 3, 5 and 7 replications.

As one will notice, the unbalanced design number 1, 4, 11 and 12 are analog to Swallow & Monaham 1984 and several other earlier studies. This allows us to see the results in a wider setting, with possibilities to easily draw comparisons with earlier works.

The actual choice of designs had three main goals:

- Spanning a broad range of practical situations.
- Focus on the situations with significance differences.
- Open for several sorts of comparisons across the different designs.

As most of these estimates show small differences for large designs, we have focused on small and medium sized designs. In the unbalanced designs, we have chosen two moderately unbalanced designs (no. 3 & 6), and two different sort of strongly unbalanced designs (no 11 & 12). In addition we have made two designs (no. 20 & 21) that are especially suitable for comparison with analog balanced designs.

For the balanced designs, we have simulated under both small (no. 1 & 3) and moderate designs (no 7 & 9), using both a moderate and small number of repetitions. In addition, one design (no. 6) has been included for comparison with the analog unbalanced designs.

During our work, we have also simulated under several other design, that have not been included in this presentation. As for the design number, there is no system or analog between balanced and unbalanced design. Unbalanced design number 1 to 12 are from Swallow & Monahan (1984), while our additional designs are given number 20 and 21. As for the balanced design, all designs are "new", and the only reason for the missing design are that the test simulations was done for a wider set of designs.

When not otherwise stated, all calculations for one design are done using the same 5000 simulated datasets.

5.1 Frequentist estimators: ML, REML, MINQUE and ANOVA

In the balanced case, we remember that Maximum Likelihood (ML) and all MINQUE and ANOVA variants give the same results. And as for the

Restricted Maximum Likelihood estimator (REML), it only takes a small shift to gain unbiasness. In practice, this leaves us with very little or no difference, so a comparison of the different frequentist estimation methods are only interesting in the unbalanced designs.

The basic study of this situation was done by Swallow & Monaham in 1984, investigating the bias and mean square error of a wide range of frequentist estimators. As we mentioned, we will start with repeating the most central designs and estimators in this study.

In this work, we have used a larger number of observations (100 000 vs. 10 000 datasets). This is done as we get a substantial uncertainty in estimating of variance of random components[1]. We have also loosened Swallow & Monahan's demand of convergence in 20 interactions, and set this limit to 50 interactions. Another difference is that we have chosen to include estimates from the datasets which have not converged. For this choice there are arguments for both approaches, and we chose the opposite of Swallow & Monahan (1984) to check out the effect of this choice.

As Swallow & Monahan (1984), we have used the "theoretical lower bound for a quadratic unbiased estimator" as a scale for the mean square errors of the σ_a^2 estimates. From section 2.1.2, we remember that in mixed effects models with normality assumptions, the unbiased quadratic estimator with lowest variance is the MIVQUE estimator using the true value as priors. Formulas for this variance can be found in Swallow & Searle (1978).

The result of the simulations are shown in table 5.3 and 5.4. Comparing this with the results in Swallow & Monahan (1984), we find that the results are very near to those of the earlier article, with no practical relevant differences.

The equivalent results from Swallow & Monaham are discussed in section 4.1. As Swallow & Monaham, we find that in most cases, all estimators perform reasonably well. One exception is for large values of σ_a^2/σ_e^2, where ANOVA and especially MINQUE(0) have a considerable larger variance than the other estimates.

In the choice between the estimators, the Maximum Likelihood estimator (ML) is generally a good choice, with a very low mean square error. Still, it also has its weaknesses. It is quite computer intensive, and gives a somewhat

[1]Looking at Swallow & Monaham 1984, we notice that the last decimals are very uncertain, and only useful for comparisons with the other methods.

σ_a^2 (with $\sigma_e^2 = 1$)		0	0.2	1	5
Design no. 1 {3,5,7}:					
ML	Bias	0.035	-0.069	-0.379	-1.748
	$\frac{\text{MSE}}{\text{QUELB}}$	0.220	0.349	0.515	0.554
REML	Bias	0.085	0.056	0.023	-0.016
	$\frac{\text{MSE}}{\text{QUELB}}$	0.796	0.872	0.969	0.995
MINUEQ(0)	Bias	0.079	0.050	0.021	-0.001
	$\frac{\text{MSE}}{\text{QUELB}}$	0.689	0.907	1.142	1.240
ANOVA	Bias	0.084	0.053	0.022	-0.007
	$\frac{\text{MSE}}{\text{QUELB}}$	0.727	0.841	0.994	1.053
QUELB		0.049	0.179	1.500	27.300
Design no. 4 {3,3,5,5,7,7}:					
ML	Bias	0.033	-0.043	-0.197	-0.863
	$\frac{\text{MSE}}{\text{QUELB}}$	0.353	0.581	0.760	0.763
REML	Bias	0.055	0.019	0.007	0.007
	$\frac{\text{MSE}}{\text{QUELB}}$	0.729	0.872	1.003	1.000
MINUEQ(0)	Bias	0.051	0.017	0.006	0.024
	$\frac{\text{MSE}}{\text{QUELB}}$	0.639	0.934	1.252	1.358
ANOVA	Bias	0.055	0.018	0.006	0.016
	$\frac{\text{MSE}}{\text{QUELB}}$	0.694	0.844	1.035	1.081
QUELB		0.018	0.070	0.599	10.900
Design no. 11 {1,1,1,1,1,1,1,19,19}:					
REML	Bias	0.072	-0.014	-0.219	-0.753
	$\frac{\text{MSE}}{\text{QUELB}}$	9.045	1.547	1.121	0.990
REML	Bias	0.146	0.108	0.031	-0.007
	$\frac{\text{MSE}}{\text{QUELB}}$	18.967	2.551	1.370	1.190
MINUEQ(0)	Bias	0.084	0.053	0.020	-0.009
	$\frac{\text{MSE}}{\text{QUELB}}$	7.075	2.409	3.324	5.258
ANOVA	Bias	0.145	0.093	0.026	-0.006
	$\frac{\text{MSE}}{\text{QUELB}}$	13.103	1.845	1.461	1.926
QUELB		0.006	0.085	0.678	8.300

Table 5.3: Frequentist estimation on unbalanced design no. 1, 4 and 11. MSE = "**M**ean **S**qaure **E**rror" and QUELB = "Theoretical lower bound for a **q**uadratic **u**nbiased **e**stimator".

σ_a^2 (with $\sigma_e^2 = 1$)		0	0.2	1	5
ML	Bias	0.017	-0.085	-0.389	-1.738
	$\frac{\text{MSE}}{\text{QUELB}}$	0.406	0.441	0.539	0.558
REML	Bias	0.057	0.039	0.020	0.009
	$\frac{\text{MSE}}{\text{QUELB}}$	2.190	1.115	1.005	1.004
MINUEQ(0)	Bias	0.038	0.020	0.010	-0.007
	$\frac{\text{MSE}}{\text{QUELB}}$	0.771	1.086	1.510	1.751
ANOVA	Bias	0.048	0.025	0.011	-0.002
	$\frac{\text{MSE}}{\text{QUELB}}$	1.032	0.909	1.178	1.344
QUELB		0.012	0.125	1.430	27.200

Table 5.4: Frequentist estimation on unbalanced design no. 12: {2,10,18} MSE = "**M**ean **S**qaure **E**rror" and QUELB = "Theoretical lower bound for a **q**uadratic **u**nbiased **e**stimator".

biased estimator. On the other hand, the REML estimator gives a theoretical unbiased estimator[2], but often gives a considerably larger mean square error.

As we have seen, these results give us clear indicators about the efficiency of the estimators, but some questions are still unanswered:

- What is the distribution of the estimates? Is it possible to assume normal distribution for confidence intervals and testing?
- How strongly are the different estimates correlated? Are some differences common under the true model, or does large differences indicate that we have the wrong model?
- How is the likelihood for small values of σ_a^2? Is it difficult to maximize in these limiting situations?

In the remainder of the section, we will try to answer these questions.

5.1.1 Distributions of σ_a^2 estimates

The distribution of the estimates are important for both confidence interval and testing theory. For large samples, the asymptomatically normal distribution theory tells us that the Maximum Likelihood estimates should follow a normal distribution, but how well does this extend to small samples?

[2]In practice, the REML estimator takes a slight bias because of truncation of negative values.

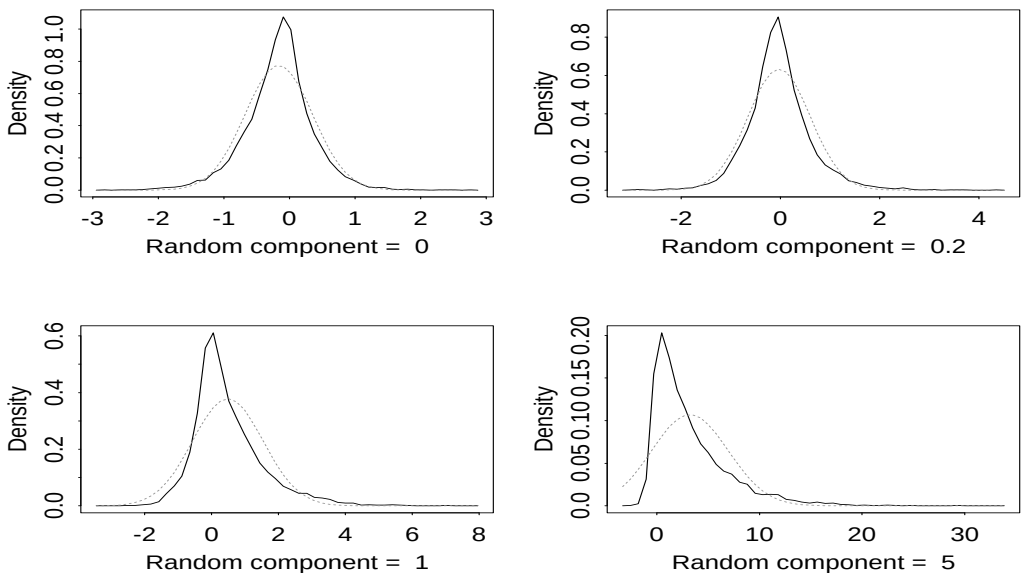


Figure 5.1: Distributions of σ_a estimate under balanced design no. 1: {2,2,2}, compared to normal distribution (dotted line). (Density is found using S-PLUS default density-function.)

In figure 5.1, 5.2, 5.3 and 5.4, we see plots of simulated[3] Maximum Likelihood estimates of the variance component (σ_a^2) for several balanced designs. These are then compared with a normal distribution with the same expectation and variance.

Looking at figure 5.1, we see that the distributions under $\sigma_a^2 = 0$ have a shape fairly near the normal distribution even with just 3 (!) individual (m) and two repetitions (n) for each individual. Still, it has somewhat large tails, something that in hypotheses testing (using the Wald test) can contribute to too many rejections of the H_0: $\sigma_a^2 = 0$ hypotheses. Comparing this to balanced design number 3 (seen in figure 5.2), we find that there is little help in increasing the number of repetitions, as the distribution even gets somewhat skewed with a small tail towards large estimates. On the other side, larger number of individuals (see figure 5.3 and 5.4) give us a very good approximation to the normal distribution.

[3]From earlier, we remember that the exact distributions in the balanced designs are known. I have, however, chosen to use simulated data, as exact calculation of the density would require large numerical intergrations. In practice, this would probably use much more processor power than simulations with the same level of precision.

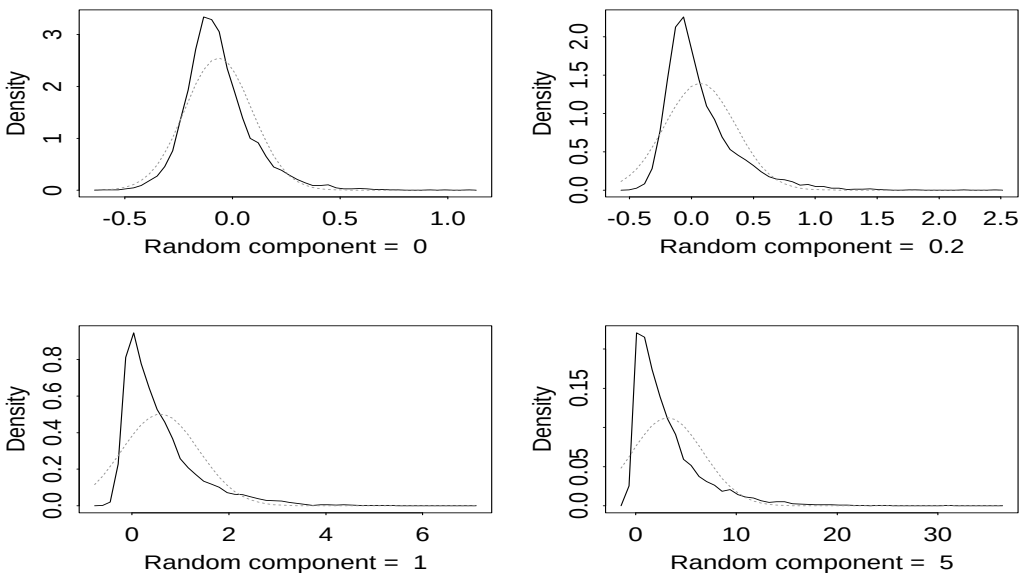


Figure 5.2: Distributions of σ_a estimate under balanced design no. 3: {5,5,5}, compared to normal distribution (dotted line). (Density is found using S-PLUS default density-function.)

The fact that the distribution sometimes is further from the normal distributions with more repetitions, can probably be attributed to the distribution being a combination of two different chi-square distributions, where the number of individual (m) and repetitions (n) give different weights for the two distributions (see section 2.2.1). We also remember from section 2.3.1, that the asymptotic theory was dependent on the number of individuals (m), not only on the total number of observations.

As σ_a^2 increases, this approximation to the normal distribution gets much less distinct. Even for $\sigma_a^2 = 0.2$, we get an asymmetrical distribution with pronounced tail to the right. Larger values of σ_a^2 further increase these differences from the normal distribution. However, looking at figure 5.3, we find that these differences from the normal distribution disappear quite quickly as the number of individual (m) increases. Looking at figure 5.1 and 5.2 versus figure 5.3 and 5.4, we can conclude that it looks like the approximation relies heavily on the number of individuals (m), while the number of repetitions (n) has only a very small impact.

Going to the unbalanced cases, we find plots of the unbalanced design number 1 ($n_i = \{3, 5, 7\}$) and 20 ($n_i = \{3, 6, 6\}$) in figure 5.5, with the cor-

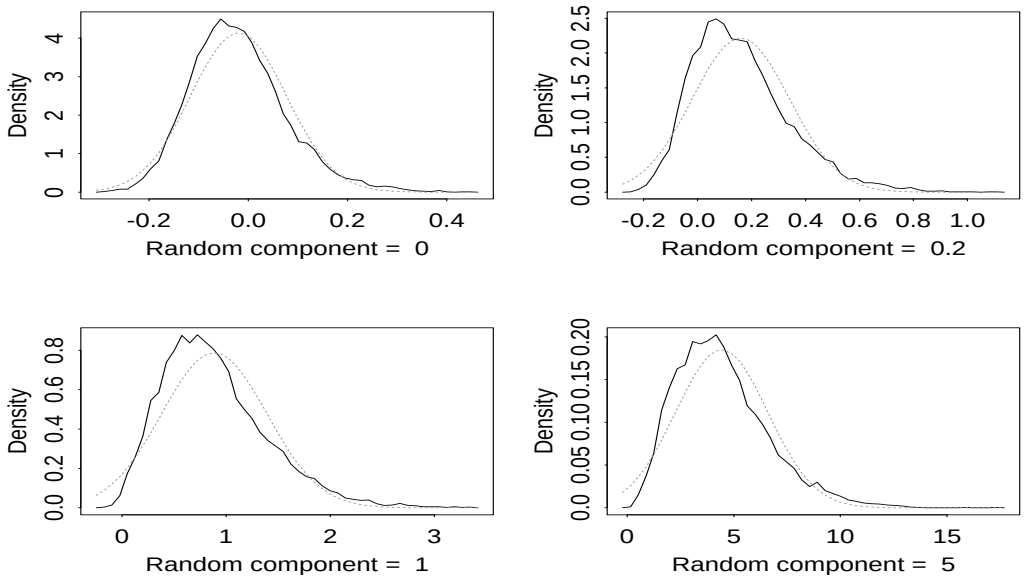


Figure 5.3: Distributions of σ_a estimate under balanced design no. 9: {5,5,5,5,5,5,5,5,5,5}, compared to normal distribution (dotted line). (Density is found using S-PLUS default density-function.)

responding balanced design (number 3: $n_i = \{5, 5, 5\}$) shown as a dotted line[4]. Here we see that for not truncated estimates, there are only moderate differences between the distributions in these balanced and moderately unbalanced design. This again means that our conclusions from the balanced designs also are valid for these moderately unbalanced designs. Looking at figure 5.6, these differences increase somewhat when the number of observations increases, but are still not very large.

Conclusions:

Based on visual examination, we generally find that the distributions are quite near the normal distribution, if we do not have very few individuals (m) combined with a moderate or large random component (σ_a^2). This gives us an indication that the asymptotical maximum likelihood theory and related methods will work even for moderate samples, something we will investigate further when looking at the Wald test in section 6.1. In general, it looks like the Wald test will have somewhat high level of rejects (of the H_0: $\sigma_a^2 = 0$

[4]For both lines, the plots are just for positive non-truncated values, but the density is scaled with the number of truncated values.

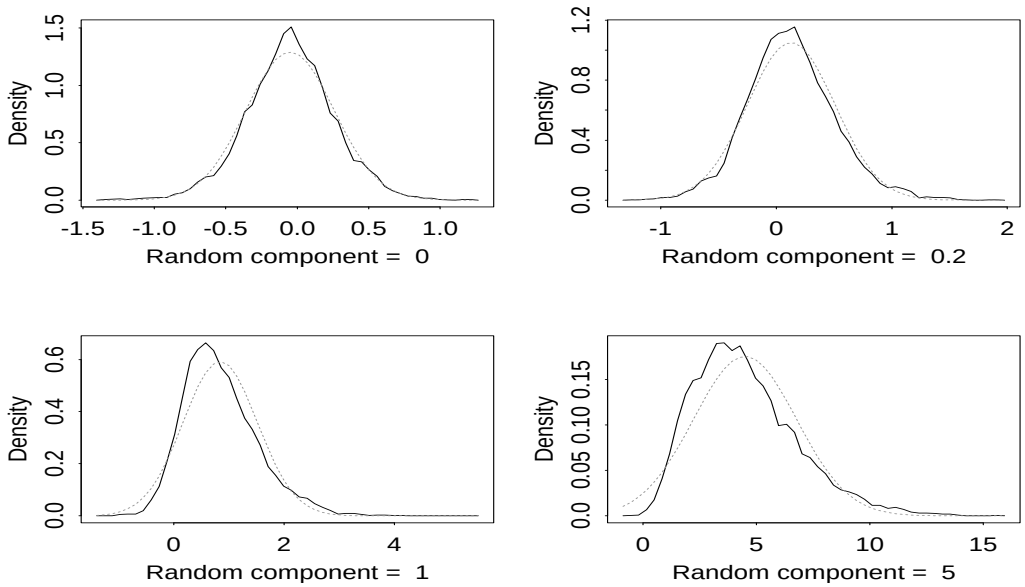


Figure 5.4: Distributions of σ_a estimate under balanced design no. 7: {2,2,2,2,2,2,2,2,2,2}, compared to normal distribution (dotted line). (Density is found using S-PLUS default density-function.)

hypotheses), while confidence intervals should work fairly well.

5.1.2 Correlations of random component (σ_a^2) estimates

In table 5.5, and figure 5.7, we find correlations between the different estimates in the unbalanced design number 4 ($n_i = \{3,3,5,5,7,7\}$). As we see, there is very good correlation between the different estimates, with an almost linear correlation between the ML (Maximum Likelihood) and REML (Restricted Maximum Likelihood) estimators.

Comparing table 5.5 and 5.6, there seems to be little influence on the correlations of both sample size or the size of the true variance component (σ_a^2). On the other hand, table 5.7, 5.8 and figure 5.8 indicate that the degree of unbalance in the model strongly affects the correlations, with a pronounced reduction in correlation as the models become more unbalanced. This should not come as a surprise, as we remember that all estimators are equal in the balanced case, except for a small bias adjustment for the REML estimate. We also notice that ML and REML are still quite close to each other even

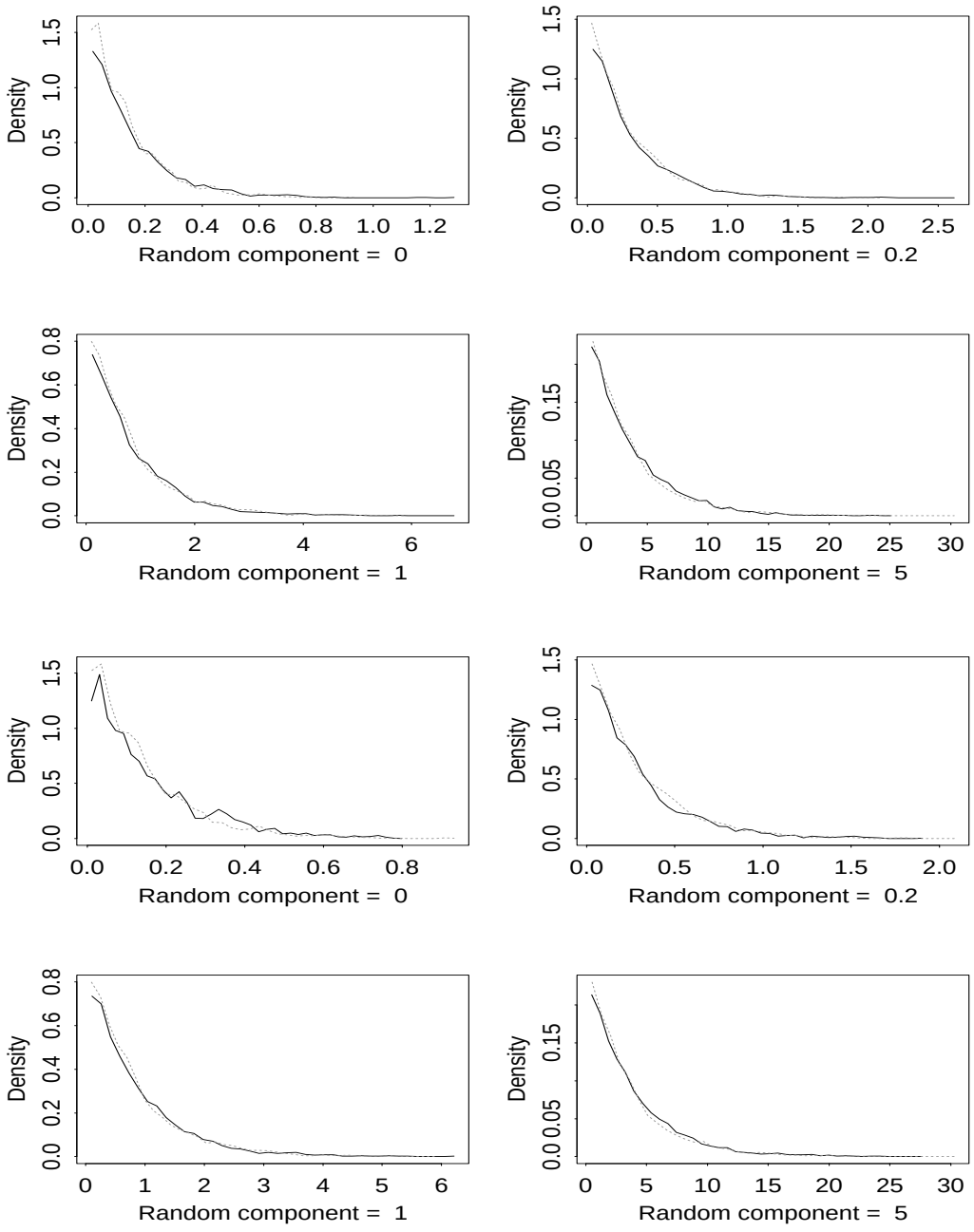


Figure 5.5: Distributions of σ_a estimate under unbalanced design no. 1: {3,5,7} (top) and no. 20: {3,6,6} (bottom), compared to balanced design number 3 (dotted line). Truncated (negative values) are taken into account, but not shown. (Density is found using S-PLUS default density-function.)

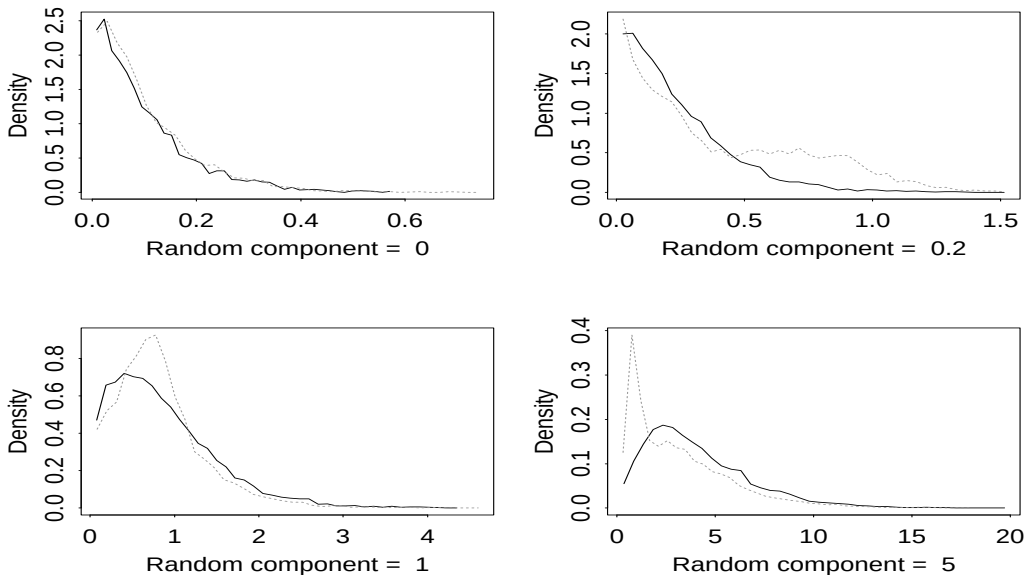


Figure 5.6: Distributions of σ_a estimate under unbalanced design no. 21: {3,3,6,6,6,6} (bottom), compared to balanced design number 6 (dotted line). Truncated (negative values) are taken into account, but not shown. (Density is found using S-PLUS default density-function.)

	ANOVA	MINQUE(O)	REML	ML
True $\sigma_a^2 = 0$				
ANOVA	1	0.92	0.98	0.96
MINQUE(O)	0.92	1	0.92	0.92
REML	0.98	0.92	1	0.98
ML	0.96	0.92	0.98	1
True $\sigma_a^2 = 1$				
ANOVA	1	0.97	0.96	0.96
MINQUE(O)	0.97	1	0.86	0.86
REML	0.96	0.86	1	$\approx$1
ML	0.96	0.86	$\approx$1	1

Table 5.5: Correlations of different estimates of σ_a under unbalanced design no. 4 : {3,3,5,5,7,7}.

	ANOVA	MINQUE(O)	REML	ML
ANOVA	1	0.97	0.98	0.98
MINQUE(O)	0.97	1	0.91	0.91
REML	0.98	0.91	1	$\approx$1
ML	0.98	0.91	$\approx$1	1

Table 5.6: Correlations of different estimates of σ_a with true value 1 under unbalanced design no. 1 : {3,5,7}.

True $\sigma_a = 0$				
	ANOVA	MINQUE(O)	REML	ML
ANOVA	1	0.36	0.80	0.65
MINQUE(O)	0.36	1	0.32	0.34
REML	0.80	0.32	1	0.85
ML	0.65	0.34	0.85	1
True $\sigma_a = 1$				
	ANOVA	MINQUE(O)	REML	ML
ANOVA	1	0.96	0.68	0.70
MINQUE(O)	0.96	1	0.46	0.49
REML	0.68	0.46	1	0.99
ML	0.70	0.49	0.99	1
True $\sigma_a = 5$				
	ANOVA	MINQUE(O)	REML	ML
ANOVA	1	0.97	0.58	0.59
MINQUE(O)	0.97	1	0.40	0.41
REML	0.58	0.40	1	$\approx$ 1
ML	0.59	0.41	$\approx$ 1	1

Table 5.7: Correlations of different estimates of σ_a under unbalanced design no. 11: {1,1,1,1,1,1,1,19,19}.

	ANOVA	MINQUE(O)	REML	ML
ANOVA	1	0.76	0.89	0.78
MINQUE(O)	0.76	1	0.55	0.58
REML	0.89	0.55	1	0.86
ML	0.78	0.58	0.86	1

Table 5.8: Correlations of different estimates of σ_a with true value 0 under unbalanced design no. 12: {2,10,18}.

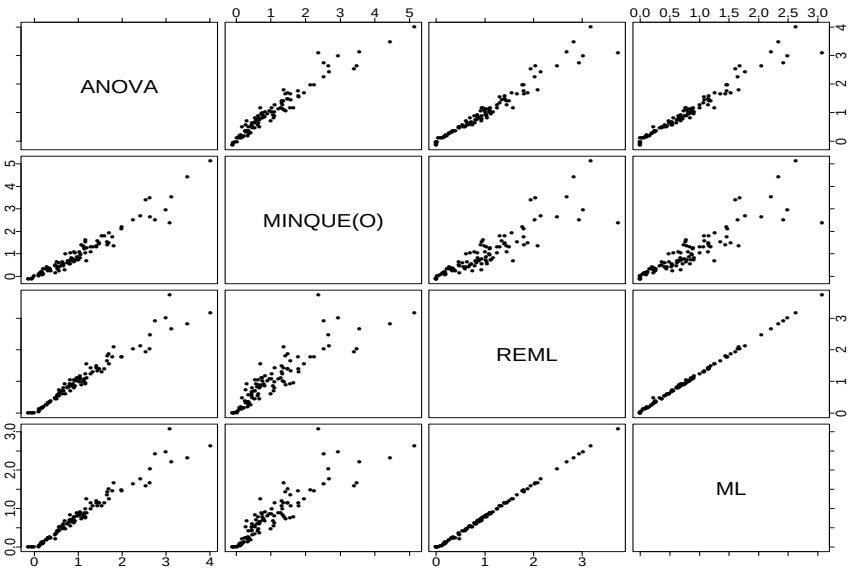


Figure 5.7: Estimates of σ_a with true value 1 under unbalanced design no. 4 : {3,3,5,5,7,7}.

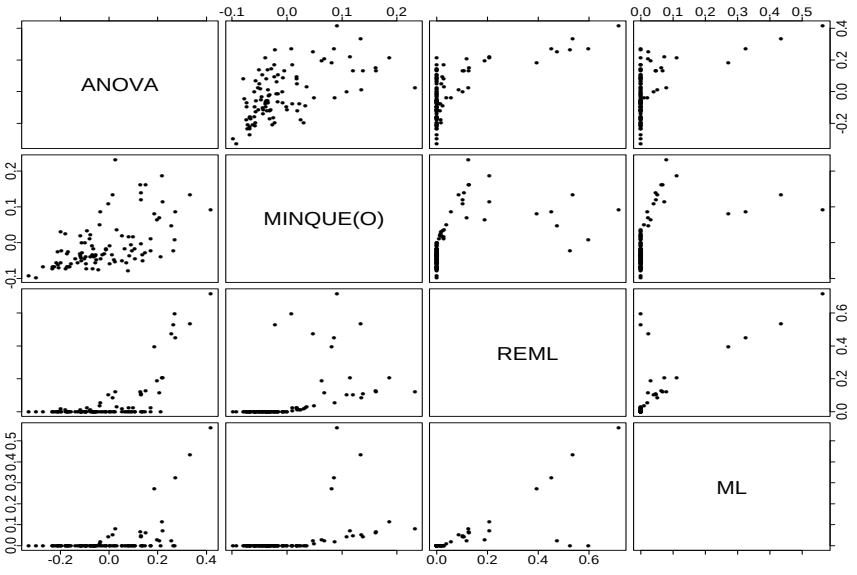


Figure 5.8: Estimates of σ_a with true value 0 under unbalanced design no. 11: {1,1,1,1,1,1,1,19,19}.

in distinctly unbalanced designs, something that probably can be attributed to their close theoretical origin.

Conclusions:

As the models become more unbalanced, the REML, ML, ANOVA and MINQUE(0) estimates can show considerable differences. These differences occur both between ANOVA - MINQUE(0), MINQUE(0) - REML/ML and ANOVA - REML/ML. The only exception is the REML and ML estimates, which in nearly all cases are very closely correlated. There are relatively small effects on the correlations as the number of observations increases. For practical work, this tells us that some differences are quite common even when we are estimating under the correct model.

5.1.3 How is the likelihood for small values of σ_a^2?

Is it hard to maximize the likelihood? Do we get limit problems for small variance components (σ_a^2)?

To throw some light on this, we have plotted the likelihood for some of the simulated designs under $\sigma_a^2 = 0$. One example of this is shown in figure

5.9. As we can see, the likelihood is easy to optimize even for an unbalanced design with small values of σ_a^2, as it has a smooth surface with just one unique top point. This is also supported by the likelihood formula, as it for this simple example is a quadratic form of the observed data.

5.2 Bayesian estimation using Gibbs sampler

As discussed in section 3.1, Bayesian estimation using the Gibbs sampler has become quite popular in recent years. Still, there have been very few studies of its performance, both absolute and relative to the standard frequentist approaches. This probably has several reasons, based on both technical, practical and methodical issues.

The main reason is probably the historical division between statisticians believing in (and using) the frequentist and Bayesian approaches. Still, there are several other problems concerning the numerical difficulties, the large number of calculations involved and problems of comparison to such different approaches. In this section, we will discuss some of these problems, and then perform a somewhat limited simulation study. In practice, it is difficult to perform a real comparison of the two methods, but we hope this study will give us some hints of the performance of Bayesian estimates, and can probably make a good foundation for further studies.

5.2.1 Philosophical issues; Comparing the different approaches

One of the first questions in comparing Bayesian and Frequentist methods, is whether these quite different approaches really can be compared to a simulation study. To answer this question, we must make a distinction between the two different situations where Bayesian statistics are used:

- When we have information in addition to the data that we want to incorporate in the model.
- When we use the Bayesian approach without any prior information, and just aim at choosing so-called "objective" or "noninformative priors". This could be done either because we like the Bayesian interpretations, or want the easy modeling capabilities of Bayesian methods, or have problems with implementation of a complex frequentist model.

  (see figure 5.10 for arguments for using Bayesian methods)

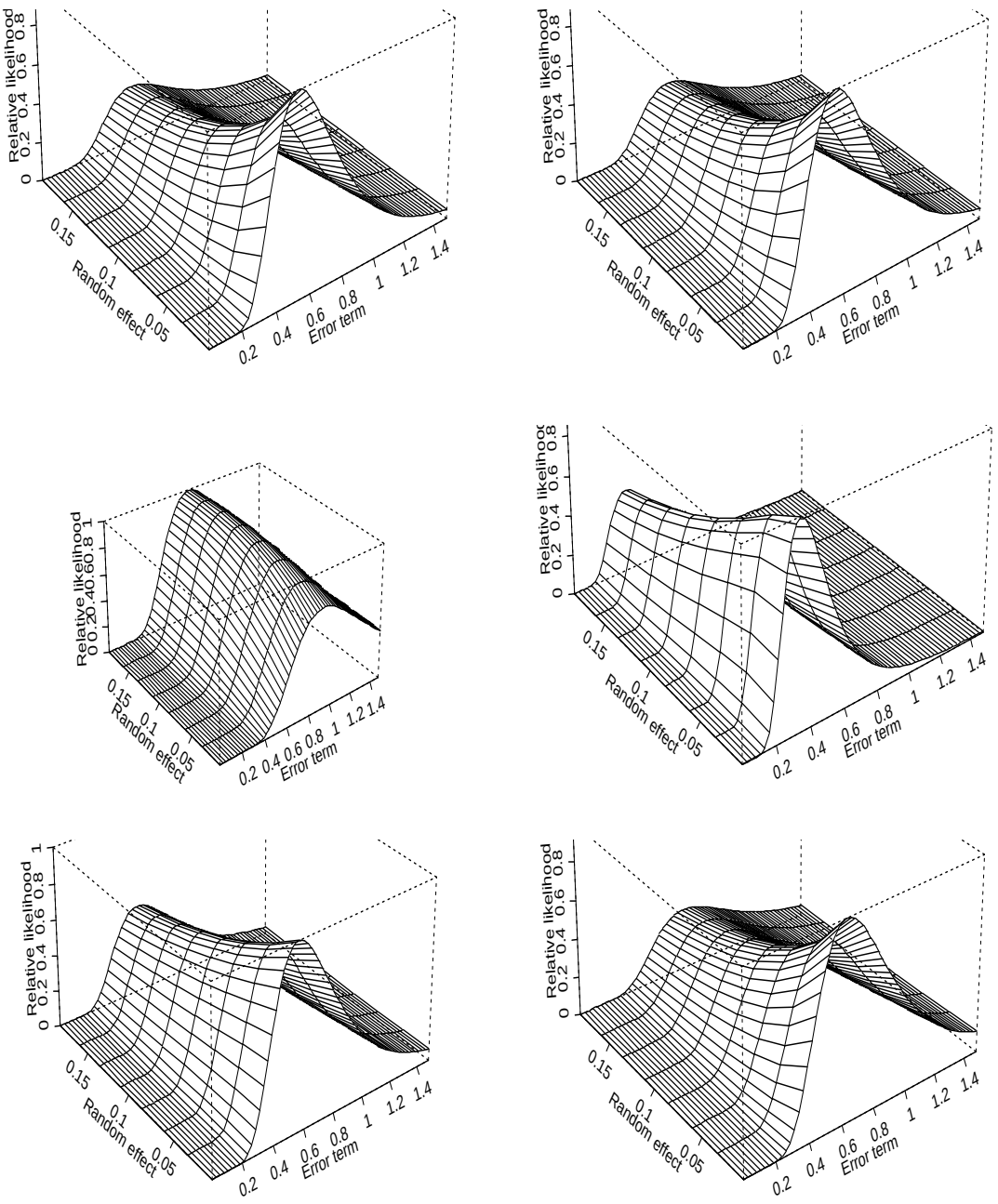


Figure 5.9: Likelihood plots for simulated data with small random components. (Design no. 1 : {3,5,7}, with $\sigma_a^2 = 0$ and $\sigma_e^2 = 1$.)

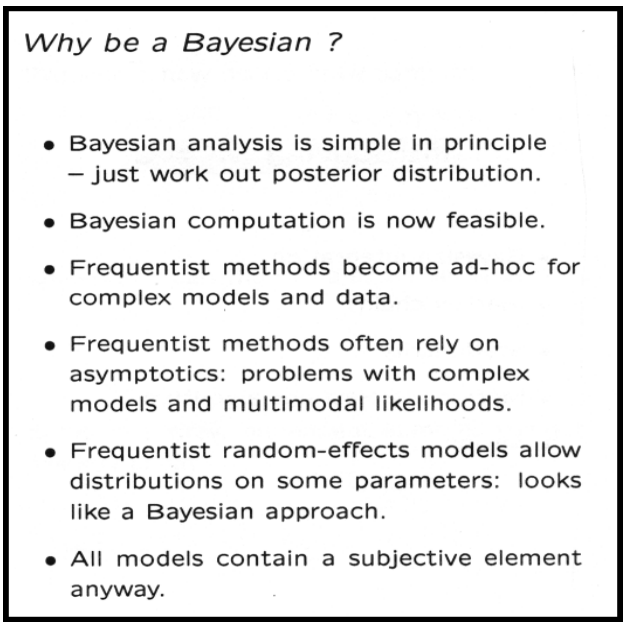


Figure 5.10: "Why be a Bayesian?" - Slide from W. R. Gilks's BUGS course, Oslo, Norway 1997.

In the first of these cases, we really have quite a different objective with our Bayesian estimate than with the traditional frequentist estimates. This means that a simulation study would only be meaningful as measurement of the impact from different priors.

In the case of "noninformative priors", we aim in some sense at the same objective, but with somewhat different interpretations of the final results. In this case, some will prefer either a Bayesian or frequentist method just because of its interpretations, but my personal belief is that users in practice are more concerned with predicting and how much they can learn from their data, than the different ways of looking at probability.

With this in mind, we will compare both approaches and see how well they are doing in predicting the true values. Still, we have many important issues about both the basis for the analysis and the final interpretations, so the results in this thesis is only one of many questions we must consider when choosing between the two different approaches.

From section 2.2.2, we also remember that except for possible numerical problems, the Bayesian methods give the correct a posteriori distribution for

the given priors. *When we have the correct answer, does it make any sense to compare the result with different methods?*

In many ways, this question has the same answer as the previous one; When we have prior information, it is interesting to study the effect of different priors, but not the efficiency of the estimate as a predictor. In the situation where we aim for "noninformative" priors, the theoretical a posteriori distribution given the priors make little practical sense, as our choice of priors are neither unique nor theoretically superior to other priors.

5.2.2 Design and limitation of the simulation study

In addition to the theoretical problems seen in the previous section, we face several other questions when performing simulation studies of the Gibbs sampler. Some of these are:

Choice of priors:

As the priors for the Gibbs sampler can have a huge impact on the estimates, we should take care in the choice of priors. Optimally, we should simulate for a wide range of priors, but in order to limit this study, we have chosen to concentrate on just two sets of priors. This is a clear limitation of the study, and other priors should be included in further studies. The choice of priors are:

"Default" prior: These priors are taken from the random effects example in the BUGS examples collection (Spiegelhalter et al. 1996). As these probably are chosen by many BUGS users, we have for convenience called it "BUGS Default" (or just "default"). This set of priors is:

$\mu \sim N(0, 1e+10)$
$\sigma_e^2 \sim \Gamma(0.001, 0.001)$ (Gamma distribution)
$\sigma_a^2 \sim \Gamma(0.001, 0.001)$ (Gamma distribution)

Remember that $\Gamma(r, \lambda)$ has an expectation of $\frac{r}{\lambda}$, and variance of $\frac{r}{\lambda^2}$. This gives us an expectation 1 and variance 1000, for both the variance component and error term priors. To some degree, this makes the priors pretty uninformable, as they span all their possible values with a reasonable expectation. We also note that this has some parallel to the choice of MINQUE(1), as both have expectation 1 for both the random and error component.

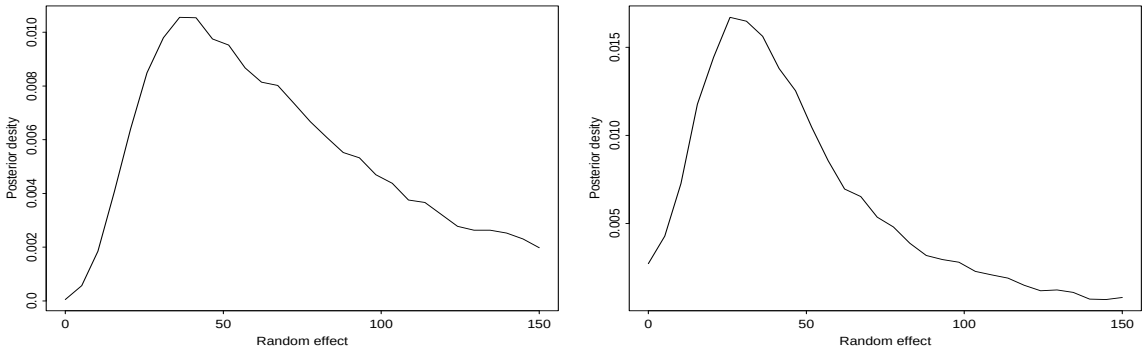


Figure 5.11: Posterior distribution of Gibbs sampler estimate on Mississippi data. (Using S-PLUS density function with n=30 and width=10.)

"Alternative prior" ("alt"-prior): In our second set of priors, we have also kept the expectation of both random and error component as 1, but with a much more moderate variation. This is done as the "default" prior with its large variation can open for some pretty extreme values, that can make a considerable impact on our estimate. These priors are called "alternative prior" (or just "alt"-prior), and is defined as:

$$\mu \sim N(0, 1e+10)$$
$$\sigma_e^2 \sim \Gamma(1,1)$$
$$\sigma_a^2 \sim \Gamma(1,1).$$

Both these prior distributions will have some limit problems under $\sigma_a^2 = 0$, as the Gibbs sampler never could take negative values and $P(X = 0) = 0$. One practical consequence of this is that the posterior distribution in practice never could take have zero mean or median. This gives the Gibbs sampler a handicap under $\sigma_a^2 = 0$, and is something that should be investigated further in future studies.

Choice of estimate from the a posteriori distribution:

From section 2.1.5, we remember that the basic Bayesian analysis just gives us a posteriori distribution. In many ordinary studies, we can just settle for this distribution, but in simulation studies, we will have to choose a point estimate to generalize the results from the vast amount of simulations. In practice, this leaves us with many possible point estimates based on the a posteriori distribution. Some of these are:

1. Posterior mean
   This is probably the most widely used method. It is implemented in BUGS, and is frequently used in the BUGS manual.

2. Posterior median
   This method is a more numerically robust alternative to the posterior mean, but also with a slightly different aim.

3. Maximum of the posterior distribution
   When the priors are uniformly distributed on the whole parameter space, this refers to the Maximum Likelihood solution. We can construct such priors for some cases, but for our random effects models we have no such priors, as the parameters can take all positive values (and no uniform distribution can span over an infinite scale).

The effect of these priors can be shown on the Mississippi river data from section 1.2. On these data, the posterior mean gives an estimated random component of 103.1 ("default") / 57.2 ("alternative"), while the posterior median gives values of 69.9 ("default") / 41.4 ("alternative"). And seen in figure 5.11, the maximum of posterior distribution for the "default"-priors come in the mid thirties, while the "alternative"-priors come in the mid twenties. From these calculations, we see that the different point estimates can give very different values, but for this study, we will limit it to the first method as it probably is the one most commonly used in practice.

Numerical problems - convergence

From section 2.1.5, we remember that the Monte Carlo integration is based on the Markov Chain coverages in distribution to its stationary distribution. Furthermore, we use a given number of observations from the stationary distribution to calculate the posteriori distribution. Using this, we have two possible traps:

1. Has the Markov Chain converged in distribution to its stationary distribution?

2. Do we have enough observations from the stationary distribution to get a good estimate of the posterior distribution?

In practice, it is recommended to check these properties by looking at the Markov Chain. For the BUGS package, this is usually done by the free S-PLUS CODA package (or the new BOA package available for both S-PLUS and R versions). Examples of CODA are shown in figure 5.11, where we

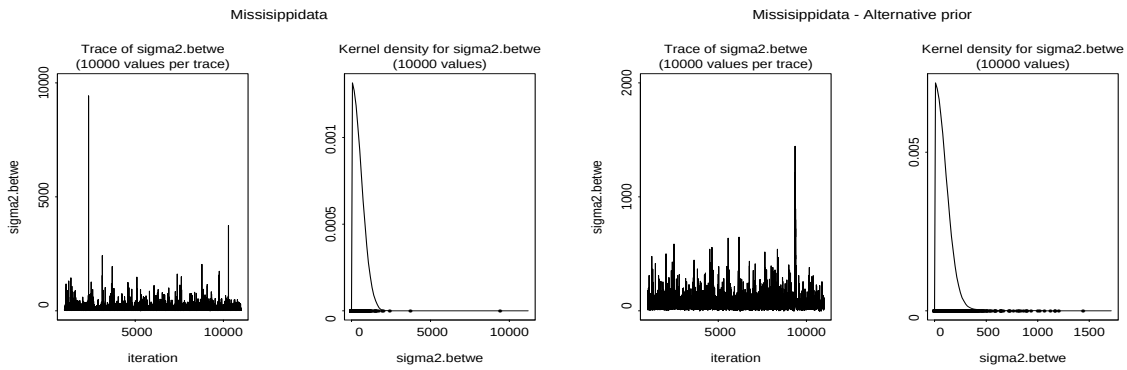


Figure 5.12: Analysis of Markov Chains (CODA output) for Gibbs sampler estimation on Mississippi river nitrogen levels data.

see an analysis of step 1001 to 11 000 of the Markov Chain estimation for the Mississippi river data from section 1.2. As we see, it looks like the first 1000 drawings (the so-called "burnin") has brought the Markov Chain to its stationary distribution. Still, there are some pretty extreme values, which indicate that we would need a large number of observations from this stationary distribution. For this, 10 000 observations are probably just enough to make a reasonable estimate.

Both these problems suggest using a quite long Markov Chain, but in practice, we must balance this against the available computer power. In most cases, the BUGS manual suggest to start with 1000 drawings, by which the Markov Chain usually converge. After that, it takes 10 000 drawings for the actual estimate. With increasingly powerful computers, these numbers often could be increased somewhat, but in a simulation study, this would be quite difficult as it involves several thousand estimates. Even at 10 000 drawings, memory problems arise and we have selected to save just one of every ten observations. This makes the estimates somewhat more unstable, but is generally much better than just using a short Markov Chain (as the sampled drawings are less dependent of each other). Small tests we have done, suggest that these numerical problems only have a limited impact, but we still have an extra uncertainty in our estimates.

Ideally, we would look at every Markov Chain to check for numerical problems, but with estimates for 5000 datasets this would be impossible in practice. For this reason, we have included not just the mean and the mean square error in the analysis of the simulation results, but also the median and the Interquartile Range (the distance between the 25 and 75 percentile).

These measurements have the advantage of being more robust for extreme values, that could arise from numerical problems.

5.2.3 Simulation results

Using the methods and choices described in the previous section, table 5.9 shows simulation results for the variance component on balanced design number 3 ($n_i = \{5, 5, 5\}$). As we see, the Gibbs sampler with "default" priors suffers from a huge variance and large bias for this very small dataset. This is most visible for the mean and mean square error, but also the more numerically robust measurements of median and interquartile range perform well below the Maximum Likelihood estimator. Generally, the Gibbs sampler with alternative priors performs much better, but still it has a long way to go in comparison with the Maximum Likelihood estimator.

In table 5.10, we see the result of increasing the number of observations to the moderate dataset of balanced design number 9 ($n_i = \{5, 5, 5, 5, 5, 5, 5, 5, 5, 5\}$). As we see, the differences quickly decrease when the number of observations increases. The Maximum Likelihood estimate is still the clear "winner" as the best predictor, but the differences is much smaller. In the choice between the two different Gibbs samplers, there are no longer a clear favorite. For small values of σ_a^2 (e.g. $\sigma_a^2 < 1$), the "default" priors give very good estimates, that are very close to the Maximum Likelihood estimate. For larger σ_a^2, the "default" priors still suffer from a notable bias and variance.

To summarize, this study indicates that even very vague "uninformative" priors have a considerable impact on small datasets. For these datasets, the Maximum Likelihood estimator is a much better predictor than the Gibbs sampler (when there is no prior information available for the construction of good priors). Still, only moderate increases in observations decreases these differences considerably, as the Gibbs sampler quickly gets properties quite close to the Maximum Likelihood estimator.

Comparing the unbalanced design number 1 ($n_i = \{3, 5, 7\}$) in table 5.11, with the balanced design number 3 ($n_i = \{5, 5, 5\}$) in table 5.9, we see that the effects of a slightly unbalanced model are very small. Overall the result almost matches the corresponding balanced model, with only a slight gain for the Maximum Likelihood and a slightly less precise "default" - Gibbs sampler. Looking at table 5.12, we see that the differences still are quite small for the strongly unbalanced model number 12 ($n_i = \{2, 10, 18\}$). As a conclusion, it looks like our results for the balanced designs also are valid for the unbalanced designs.

Regarding the correlations between the two different methods, we find in figure 5.13 plots for the 100 first estimates of balanced design number 1 ($n_i = \{5, 5, 5\}$) with $\sigma_a = 1$. As we see, there are very good correlations between the Maximum Likelihood estimator and Gibbs sampler with alternative priors. On the other hand, there is only a moderate correlation between these estimates and the Gibbs sampler with "default" priors.

For small variance components (see figure 5.14), the correlation almost vanishes, while the degree of unbalance in the model seems to have little effect (figure 5.15 and 5.18).

The most distinct effect seems to come from the number of observations, and looking at figure 5.16 we see that a large number of observations gives us almost a linear correlation between the different estimates. Combined with small variance components (see figure 5.17), these correlations get less distinct, but are still visible.

Conclusions:

When the number of observations increases, the connection between the Gibbs sampler and Maximum Likelihood estimate increases to an almost linear correlation. This occurs already on quite moderate datasets if the variance component is not too small. In small datasets, the correlations varies quite a lot with the different choices of priors, and the Gibbs sampler estimates are often far from the Maximum Likelihood estimate. As for the degree of unbalance in the underlying design, the effect is very small both on the estimates and the correlations between the different estimates.

Remark: In contrast to the earlier study of frequentist methods, negative Maximum Likelihood estimates are not truncated in this study. This is done in order to see clearer what happens in the correlation plots.

σ_a^2	Bias of mean			Bias of median		
	Std	Alt	ML	Std	Alt	ML
0	1.07	1.29	-0.07	0.48	1.24	-0.09
0.2	1.87	1.24	-0.13	0.59	1.14	-0.21
1	7.03	1.05	-0.41	1.79	0.70	-0.65
5	29.79	0.69	-1.81	11.58	-1.06	-2.91
σ_a^2	Mean square error			Interquartile Range		
	Std	Alt	ML	Std	Alt	ML
0	21.27	1.73	0.03	0.57	0.20	0.17
0.2	54.04	1.69	0.10	1.49	0.35	0.30
1	1865.94	2.17	0.80	6.43	0.95	0.83
5	16955.38	30.19	15.85	29.11	4.98	3.56

Table 5.9: Gibbs sampler estimation on balanced design no. 3: {5,5,5}. Std = Gibbs sampler with "stadard" priors, Alt = Gibbs sampler with "alternative" priors and ML = Maximum Likelihood estimate (for comparison).

σ_a^2	Bias of mean			Bias of median		
	Std	Alt	ML	Std	Alt	ML
0	0.07	0.44	-0.02	0.05	0.43	-0.03
0.2	0.03	0.37	-0.04	-0.05	0.34	-0.07
1	0.24	0.24	-0.11	0.14	0.14	-0.19
5	1.34	0.09	-0.56	0.92	-0.24	-0.85
σ_a^2	Mean square error			Interquartile Range		
	Std	Alt	ML	Std	Alt	ML
0	0.01	0.20	0.01	0.04	0.07	0.12
0.2	0.05	0.16	0.03	0.23	0.17	0.23
1	0.61	0.32	0.27	0.98	0.64	0.66
5	11.34	5.73	4.99	3.98	3.07	2.79

Table 5.10: Gibbs sampler estimation on balanced design no. 9: {5,5,5,5,5,5,5,5,5}. Std = Gibbs sampler with "stadard" priors, Alt = Gibbs sampler with "alternative" priors and ML = Maximum Likelihood estimate (for comparison).

σ_a^2	Bias of mean			Bias of median		
	Std	Alt	ML	Std	Alt	ML
0	1.13	1.31	0.04	0.51	1.26	0.00
0.2	1.92	1.25	-0.07	0.64	1.16	-0.20
1	5.84	1.06	-0.40	1.66	0.70	-0.68
5	29.84	0.71	-1.74	12.11	-0.95	-2.78
σ_a^2	Mean square error			Interquartile Range		
	Std	Alt	ML	Std	Alt	ML
0	27.52	1.79	0.01	0.64	0.21	0.00
0.2	91.38	1.72	0.06	1.46	0.34	0.17
1	435.09	2.67	0.78	6.22	0.96	0.87
5	42500.58	26.18	14.68	30.13	5.33	3.91

Table 5.11: Gibbs sampler estimation on unbalanced design no. 1: {3,5,7}. Std = Gibbs sampler with "stadard" priors, Alt = Gibbs sampler with "alternative" priors and ML = Maximum Likelihood estimate (for comparison).

σ_a^2	Bias of mean			Bias of median		
	Std	Alt	ML	Std	Alt	ML
0	0.82	1.27	0.02	0.37	1.21	0.00
0.2	1.75	1.21	-0.09	0.54	1.12	-0.20
1	8.08	1.10	-0.36	2.19	0.74	-0.64
5	27.29	0.64	-1.81	11.15	-1.07	-2.86
σ_a^2	Mean square error			Interquartile Range		
	Std	Alt	ML	Std	Alt	ML
0	25.40	1.72	0.01	0.44	0.19	0.00
0.2	218.75	1.68	0.05	1.44	0.32	0.15
1	6982.55	2.53	0.81	6.98	1.04	0.90
5	12070.89	27.00	14.98	27.14	5.05	3.76

Table 5.12: Gibbs sampler estimation on unbalanced design no. 12: {2,10,18}. Std = Gibbs sampler with "stadard" priors, Alt = Gibbs sampler with "alternative" priors and ML = Maximum Likelihood estimate (for comparison).

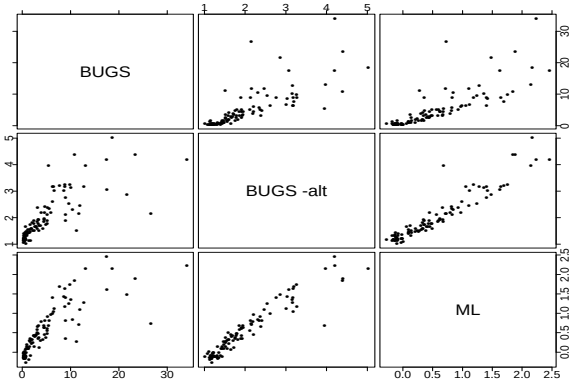


Figure 5.13: BUGS vs. ML on balanced design no. 3: {5,5,5} with $\sigma_a^2 = 1$.

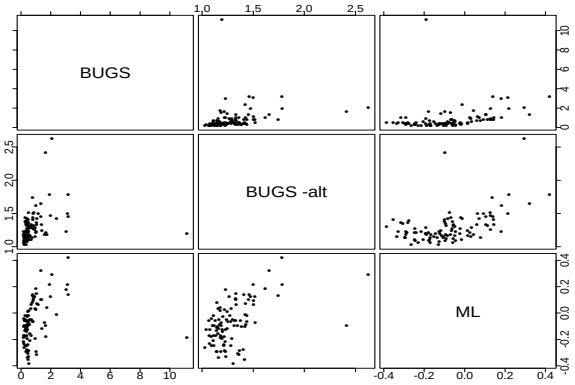


Figure 5.14: BUGS vs. ML on balanced design no. 3: {5,5,5} with $\sigma_a^2 = 0$.

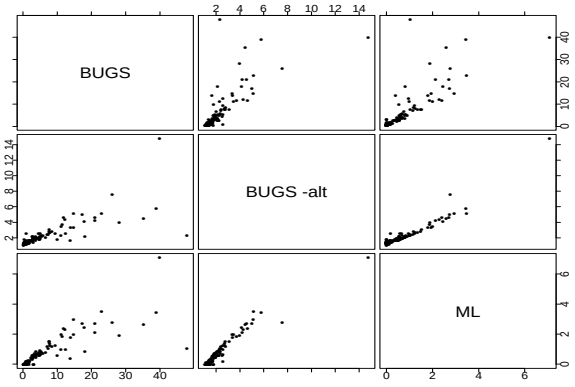


Figure 5.15: BUGS vs. ML on unbalanced design no. 1: {3,5,7} with $\sigma_a^2 = 1$.

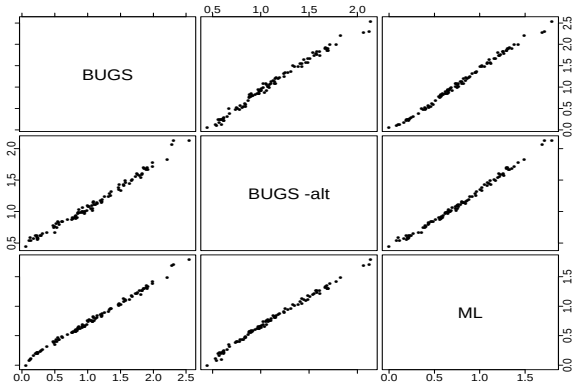


Figure 5.16: BUGS vs. ML on balanced design no. 9: {5,5,5,5,5,5,5,5,5,5} with $\sigma_a^2 = 1$.

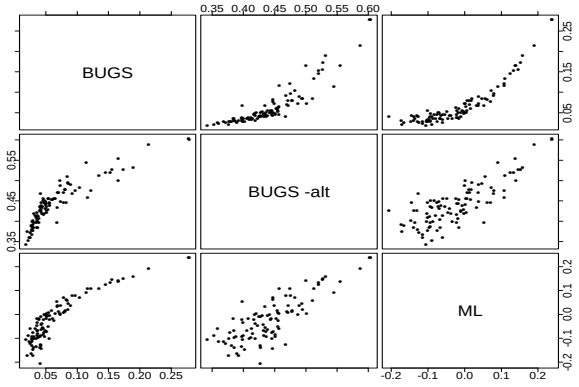


Figure 5.17: BUGS vs. ML on balanced design no. 9: {5,5,5,5,5,5,5,5,5,5} with $\sigma_a^2 = 0$.

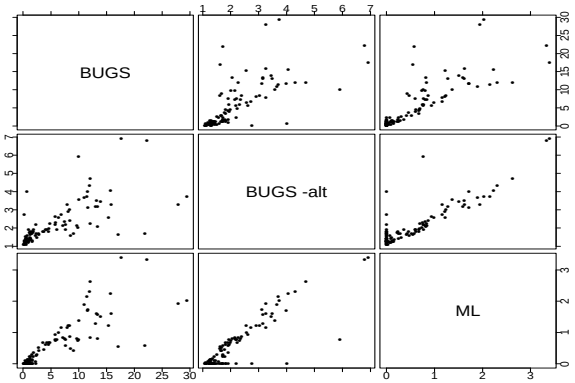


Figure 5.18: BUGS vs. ML on unbalanced design no. 12: {2,10,18} with $\sigma_a^2 = 1$.

Chapter 6

A Monte Carlo study of different testing methods

> A beautiful theory, killed by a nasty, ugly little fact!
> Thomas Henry Huxley (1825-1895)

As we saw in section 2.3, there are several different testing methods in mixed effects models. This leaves two questions for the applied statistician:

- Does the choice of testing method really matter?
- And if yes; What methods are recommended for the different situations?

Today, SAS is one of the leading statistical packages. As one of the first to include mixed effects models, its Proc Mixed routines have become one of the most used packages for analyzing mixed effects data. In figure 6.1, we find an output from the SAS Proc Mixed routine performed on the Mississippi river data (from section 1.2).

As we can see, the SAS runs a Wald test that gives us a p-value of 0.08[1] (e.g. 8 %) for the hypothesis of non variance component (E.g. H_0: $\sigma_a^2 = 0$). This is clearly not significant on any reasonable level, but doing the same test using Likelihood ratio theory, we find a p-value of 0.00018 (e.g. 0.0 18%). So:

- Why do we see such large differences?

[1] As random effects only can take positive values, H_0: $\sigma_a^2 = 0$ should only be tested as a one-sided test (E.g. H_a: $\sigma_a^2 \geqslant 0$). In spite of this, SAS uses its standard two sided p-value ($\Pr > |Z|$), and users must in practice divide this p-value by 2.

Design number	Number of observations
10	$m = 5$, $n_i = \{10, 10, 10, 10, 10\}$
17	$m = 30$, $n_i = \{2, 2, 2, 2, 2, \ldots, 2, 2\}$
20	$m = 30$, $n_i = \{5, 5, 5, 5, 5, 5, \ldots, 5, 5\}$

Table 6.1: Additional balanced designs for the study of different testing methods. For the other balanced designs, see chapter 5.

- What are the "true p-values" under H_0 for the different methods?
- Which methods give the best power under H_1?
- And ultimately; Which tests are recommended for practical use?

In this chapter, we will try to answer these questions both by simulations and some theoretical considerations. During this work, we will focus on the balanced one-way design, but this study has probably also some relevance for unbalanced designs. Still, this is a question that must be investigated further in new studies.

Design of study

For the study of different testing methods, we have used balanced design number 1, 3, 7 and 9 described in section 5. These designs are chosen as they represent small and medium designs, with a small and moderate number of replications. In addition to these designs, I have also included three new designs seen in figure 6.1. These designs are added so that we see what happens with a somewhat larger number of individuals or replications. As for the simulation of estimating methods in chapter 5, we have chosen to fix μ as 3 and σ_e^2 as 1. Under this we have simulated under H_0: $\sigma_a^2 = 0$ and different values of H_1 : $\sigma_a^2 = \{0.1, 0.2, 1, 5, 25\}$.

As for the Wald test, exact distributions of the quadratic forms are known. In theory, we could use this to evaluate the Wald-test through theoretical work, by deducing the exact distribution of $\hat{\sigma}_a^2$. This would, however, in practice include huge algebra, that probably would lead to such large equations that it would be very hard or impossible to use it in any direct evaluation of the practical consequences. For this reason, we would probably end up needing numerical tables anyway. In addition, the simulations also give us direct comparisons with other methods, as all tests are done on the same datasets. Also for a comparison with the other methods, the results of the

Output 4.1 *The Results of Using PROC MIXED to Obtain REML Estimates of the Variance Components and to Obtain Estimated BLUP for Each INFLUENT*

Covariance Parameter Estimates (REML)

Cov Parm	Ratio	Estimate	Std Error	Z	Pr > |z|
INFLUENT	1.48438019	63.32114984	45.23148440	1.40	0.1615
Residual	1.00000000	42.65830963	10.85707766	3.93	0.0001

Figure 6.1: Exsample of SAS Proc Mixed routine on the Mississippi river data (see section 1.2) from the book "SAS System for Mixed Models" (Littell et al. 1996).

F-test under H_0 given (even we know from section 2.3.3, that the F-test gives exact significant levels).

6.1 Simulation of different testing methods

In table 6.2 to 6.8, we find simulation results for the different testing methods; three variants of the Wald test, the Likelihood ratio test and the F-test. For more information on these tests, see section 2.3.

The tables show the proportions of tests rejected for different significant levels (5 % and 1 %), simulated under H_0 ($\sigma_a^2 = 0$) and different values of H_1 ($\sigma_a^2 = \{0.1, 0.2, 1, 5, 25\}$). The optimal test would reject H_0 for all tests performed under H_1, while still never rejecting H_0 when it is correct (e.g. $\sigma_a^2 = 0$). In practice, there is no such test, and we aim at a test that has the highest proportion of rejections under H_1 (called power), while still not rejecting more tests under H_0 than the chosen significant level.

This leads us to two criteria for evaluation of the different testing methods used in mixed models:

1. Do the tests keep the limit of rejected tests under H_0, set by the significant level?

2. How is the test's power under H_1? (E.g. proportions of rejected tests under H_1.)

The results for the different tests may be summarized in the following conclusions:

		σ_a^2					
	Level	H0	.1	.2	1	5	25
F	0.05	0.048	0.064	0.072	0.177	0.514	0.837
	0.01	0.008	0.014	0.018	0.042	0.206	0.609
LR	0.05	0.019	0.028	0.036	0.079	0.327	0.720
	0.01	0.003	0.005	0.010	0.022	0.121	0.461
Wald	0.05	0.000	0.000	0.000	0.000	0.000	0.000
(REML)	0.01	0.000	0.000	0.000	0.000	0.000	0.000
Wald	0.05	0.000	0.000	0.000	0.000	0.000	0.000
(ML)	0.01	0.000	0.000	0.000	0.000	0.000	0.000
Wald	0.05	0.212	0.260	0.287	0.495	0.799	0.948
(H_0)	0.01	0.172	0.212	0.236	0.432	0.758	0.937

Table 6.2: Proportion of H_0: $\sigma_a^2 = 0$ tests rejected for balanced design no. 1: {2,2,2}.

		σ_a^2					
	Level	H_0	.1	.2	1	5	25
F	0.05	0.050	0.119	0.188	0.537	0.855	0.970
	0.01	0.012	0.032	0.074	0.337	0.757	0.945
LR	0.05	0.010	0.028	0.068	0.320	0.743	0.943
	0.01	0.002	0.008	0.023	0.181	0.640	0.914
Wald	0.05	0.000	0.000	0.000	0.000	0.000	0.000
(REML)	0.01	0.000	0.000	0.000	0.000	0.000	0.000
Wald	0.05	0.000	0.000	0.000	0.000	0.000	0.000
(ML)	0.01	0.000	0.000	0.000	0.000	0.000	0.000
Wald	0.05	0.114	0.217	0.304	0.649	0.900	0.979
(H_0)	0.01	0.071	0.151	0.234	0.583	0.876	0.975

Table 6.3: Proportion of H_0: $\sigma_a^2 = 0$ tests rejected for balanced design no. 3: {5,5,5}.

		σ_a^2					
	Level	H0	.1	.2	1	5	25
F	0.05	0.051	0.089	0.125	0.481	0.971	1.000
	0.01	0.010	0.020	0.037	0.219	0.882	0.999
LR	0.05	0.025	0.040	0.066	0.328	0.937	1.000
	0.01	0.004	0.010	0.017	0.129	0.801	0.998
Wald	0.05	0.016	0.029	0.048	0.270	0.912	1.000
(REML)	0.01	0.000	0.000	0.000	0.000	0.000	0.000
Wald	0.05	0.021	0.035	0.056	0.298	0.926	1.000
(ML)	0.01	0.000	0.000	0.000	0.000	0.000	0.000
Wald	0.05	0.192	0.276	0.352	0.783	0.996	1.000
(H_0)	0.01	0.132	0.204	0.265	0.694	0.992	1.000

Table 6.4: Proportion of H_0: $\sigma_a^2 = 0$ tests rejected for balanced design no. 7: {2,2,2,2,2,2,2,2,2,2}.

		σ_a^2					
	Level	H_0	.1	.2	1	5	25
F	0.05	0.050	0.292	0.494	0.913	0.992	1.000
	0.01	0.009	0.129	0.300	0.838	0.986	0.999
LR	0.05	0.009	0.122	0.289	0.831	0.986	0.999
	0.01	0.002	0.049	0.159	0.739	0.980	0.998
Wald	0.05	0.000	0.004	0.033	0.522	0.954	0.997
(REML)	0.01	0.000	0.000	0.000	0.000	0.000	0.000
Wald	0.05	0.000	0.006	0.041	0.554	0.959	0.997
(ML)	0.01	0.000	0.000	0.000	0.000	0.000	0.000
Wald	0.05	0.041	0.260	0.462	0.905	0.991	1.000
(H_0)	0.01	0.016	0.162	0.345	0.858	0.988	1.000

Table 6.5: Proportion of H_0: $\sigma_a^2 = 0$ tests rejected for balanced design no. 10: {10,10,10,10,10}.

		σ_a^2					
	Level	H0	.1	.2	1	5	25
F	0.05	0.051	0.214	0.406	0.952	1.000	1.000
	0.01	0.012	0.076	0.207	0.883	1.000	1.000
LR	0.05	0.016	0.092	0.237	0.898	1.000	1.000
	0.01	0.002	0.028	0.105	0.798	0.999	1.000
Wald	0.05	0.001	0.008	0.043	0.675	0.996	1.000
(REML)	0.01	0.000	0.000	0.000	0.000	0.000	0.000
Wald	0.05	0.001	0.013	0.057	0.711	0.997	1.000
(ML)	0.01	0.000	0.000	0.000	0.000	0.000	0.000
Wald	0.05	0.103	0.341	0.539	0.971	1.000	1.000
(H_0)	0.01	0.055	0.223	0.418	0.955	1.000	1.000

Table 6.6: Proportion of H_0: $\sigma_a^2 = 0$ tests rejected for balanced design no. 9: {5,5,5,5,5,5,5,5,5,5}.

		σ_a^2					
	Level	H_0	.1	.2	1	5	25
F	0.05	0.048	0.125	0.225	0.907	1.000	1.000
	0.01	0.009	0.035	0.073	0.732	0.999	1.000
LR	0.05	0.017	0.064	0.130	0.828	1.000	1.000
	0.01	0.005	0.016	0.037	0.612	0.999	1.000
Wald	0.05	0.061	0.152	0.266	0.927	1.000	1.000
(REML)	0.01	0.006	0.024	0.048	0.662	0.999	1.000
Wald	0.05	0.056	0.143	0.250	0.920	1.000	1.000
(ML)	0.01	0.006	0.022	0.046	0.656	0.999	1.000
Wald	0.05	0.161	0.315	0.478	0.979	1.000	1.000
(H_0)	0.01	0.097	0.213	0.355	0.954	1.000	1.000

Table 6.7: Proportion of H_0: $\sigma_a^2 = 0$ tests rejected for balanced design no. 17. (30 individuals with 2 observations each.)

		σ_a^2					
	Level	H_0	.1	.2	1	5	25
F	0.05	0.053	0.417	0.776	1.000	1.000	1.000
	0.01	0.013	0.203	0.567	1.000	1.000	1.000
LR	0.05	0.020	0.262	0.641	1.000	1.000	1.000
	0.01	0.003	0.115	0.419	1.000	1.000	1.000
Wald	0.05	0.110	0.569	0.870	1.000	1.000	1.000
(REML)	0.01	0.017	0.243	0.619	1.000	1.000	1.000
Wald	0.05	0.105	0.549	0.861	1.000	1.000	1.000
(ML)	0.01	0.017	0.239	0.611	1.000	1.000	1.000
Wald	0.05	0.214	0.723	0.937	1.000	1.000	1.000
(H_0)	0.01	0.153	0.641	0.904	1.000	1.000	1.000

Table 6.8: Proportion of H_0: $\sigma_a^2 = 0$ tests rejected for balanced design no. 20. (30 individuals with 5 observations each.)

F-test

From section 2.3.3, we know that the F-test is based on an exact distribution. This gives us an exact p-value under H_0, so the most interesting part in this simulation study becomes the power under H_1. Looking at table 6.3, we see that the F-test performs quite well. As shown in table 6.6 and 6.8, this is especially true when the number of observations increases. A comparison of table 6.2, 6.3 and 6.4, shows that this applies to both increases in the number of observations and individuals.

Likelihood ratio (LR) - test

Loking at table 6.3, we find that the Likelihood ratio test is fairly good, but quite conservative with significant levels several times under the correct value. From the larger designs of table 6.5, 6.7 and 6.8, we find that this is not solved either by increasing the number of individual (m) or repetitions (n). With this exception, the test performs very well with performance close to the analog F test with equal "real significant level". Anyway, we will in practice not have the possibility to calibrate the Likelihood ratio test, so that the F-test will in practical applications give us somewhat better power (under H_1). Still, there are many cases where there is no precise F-test available, and in these cases there is probably not very much to gain from using an approximate F-test.

Remark: The Likelihood ratio test is based on standard Maximum Likelihood estimation (not REML), and we have allowed rejection only for positive values of σ_a^2 (e.g.: All tests with negative σ_a^2 estimates have not been rejected).

Wald test

In the tables, we find three different variants of the Wald test. For all these tests, we have chosen to base the variance estimates on the asymptotic value. This is done because it is implemented in several statistical packages, and is probably the most commonly used approach today.

The first two variants of the Wald test are the "classical" methods for the ML (Maximum Likelihood) and REML (Restriced Maximum Likelihood) estimates. These are named Wald-REML and Wald-ML.

In addition to these two "classical" methods, we have introduced a third approach: Note that $\sigma_a^2 = 0$ under H_0. For some designs, this can be used in the estimating the standard deviation of the σ_a^2 estimate. One example of this is in the one-way balanced design, where we have the formulas for the standard deviation and easily can set $\sigma_a^2 = 0$. (Remark: To my knowledge, this has no elementary extension to the unbalanced case.)

Looking at the simulation results, the probably most striking result is that both the standard methods of the Wald test often never rejects the H_0 hypothesis, even for quite large differences. For the very small designs of balanced design number 1 and 3 (see table 6.2 and 6.6), this happens both at 5% and 1% significant level. While for the somewhat larger designs of balanced design 7, 9 and 10 (see table 6.4, 6.5 and 6.6), it happens only for the 1% significant level. Also, for 5% significant level in balanced design number 9, these Wald tests are very conservative with a true significant level just above zero. This gives the tests very low significant levels under H_0, but with equally low power under H_1. With such a low power, both these tests are in practice useless for these quite small designs. In some cases, they can even do more harm than good, as they often do not reject H_0 for data that at a first look really look as a clear rejection of H_0. As illustrated in figure 6.2, these tests have particularly trouble with low significant levels.

Increasing the number of observations, both versions of the standard Wald test give much better power. At first look, this seems very good (see balanced design number 17 in table 6.7), but looking at table 6.8, the test also sometimes gives too high significant levels (under H_0). In practice, this would be

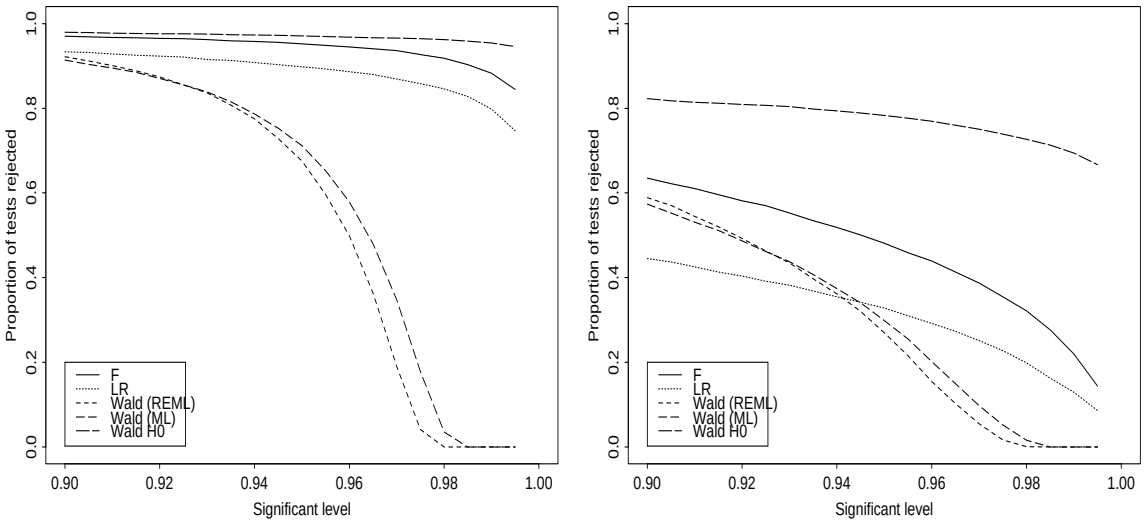


Figure 6.2: Proportion of tests rejected at different significant levels for design no. 7 : {2,2,2,2,2,2,2,2,2,2} (left) and no. 9 : {5,5,5,5,5,5,5,5,5,5} (right) and $\sigma_a^2 = 1$. Scale $= 1 - \alpha$.

quite dangerous, as it leads to too large probability of rejecting H_0 (E.g. too large probability of type I error).

All in all, the standard Wald test gives quite uncertain results. Sometimes it will never reject H_0, while in some other designs it gives significant levels well over the correct value. For practical purposes, this means that the test is quite dangerous, and could not be recommend for use in small and medium sized designs.

As for the new third approach, The "Wald - H_0", this test gives us totally different results. It has very good power under H_1 for all designs, but except for design number 10, it also gives us way too large significant leves (under H_0). As this leads to all too large probability of rejecting H_0, this test can not be recommended for any practical purposes.

6.2 Theoretical considerations

In the preceding section, we have seen considerable problems with some commonly used testing methods. At first look, these results may seem quite strange, and in this section, we will present some theoretical considerations that can shed some light on what goes wrong with these methods.

6.2.1 Why does the Wald test give so uncertain results?

As seen in section 6.1, the Wald test will for some designs never reject H_0, while it for other designs gives too high significance levels. Comparing table 6.4 and 6.6, it even sometimes performs worse with more observations. As shown below, this probably is a combination of several problems concerning use of the Wald test for random components.

Why does the Wald test for some designs never reject H_0?

In most cases of fixed effects, the standard Wald test performs fairly well, but what goes wrong with the theory in the case of random components?

To answer this question, we must go back to our assumptions for the Wald test. The most commonly discussed assumptions in connection with fixed effects are:

1. The estimator should take an approximately normal distribution.
2. We should have a fairly good estimate of the standard error.

Looking at the first of these assumptions, we remember that we discussed the distribution in section 5.1.1. Here, we sometimes found a considerable deviance from the normal distribution when we had a small number of individuals (m) combined with a moderate or large variance component. Still, under H_0, these deviancies from the normal distribution, with a tail towards large values, should give too many, not too few, rejections of H_0. Comparing the test results with the distributions found in section 5.1.1, we also find that the Wald test sometimes never rejects H_0 even when the estimates distributions are quite close to the normal distribution. We could also suspect the standard error estimate, but as shown in section 2.3.1, the asymptotical theory gives a very good estimate of the true value in the balanced one-way design.

Are there any more assumptions behind the Wald test that could go wrong?

Going back to the asymptotical Wald-test theory, we find one more assumption; The standard error estimate should be independent of the estimator. This is no problem for fixed effects, but what happens when we use this theory on random components?

Using the variance estimate from section 2.3.1, we find that the Wald test statistic is

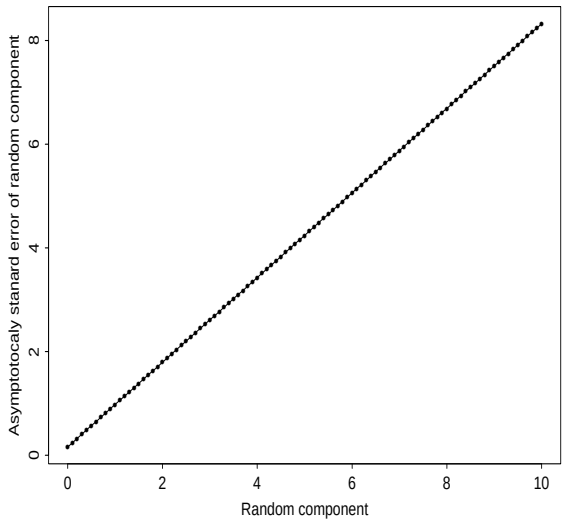


Figure 6.3: Random component versus asymptotical standard error of random component under balanced design nr 3.

$$\frac{\widehat{\sigma}_a^2}{\sqrt{2\left(\widehat{\sigma}_a^4 + 2\frac{\widehat{\sigma}_e^2\widehat{\sigma}_a^2}{n} + \frac{\widehat{\sigma}_e^4}{n^2}\right)/m}}.$$

In this formula, we find $\widehat{\sigma}_a^2$ in both the numerator and denominator, as the standard deviation estimate depends heavily on $\widehat{\sigma}_a^2$. As shown in figure 6.3, we see that this dependence in practice can be almost linear. This gives much larger estimated standard error for large estimates. In practice, this makes it very hard to get tests rejected, as the standard error of the estimate more or less eats up the gain of a large estimate.

This chould be illustrated algebraically, by letting $\widehat{\sigma}_e^2 = x$ in the test observator equation:

$$z = \frac{x}{\sqrt{2\dfrac{x^2 + 2\dfrac{x}{n} + \dfrac{1}{n^2}}{m}}},$$

and solving the equation with regards to x. This gives us one (possible) solution for positive values of z:

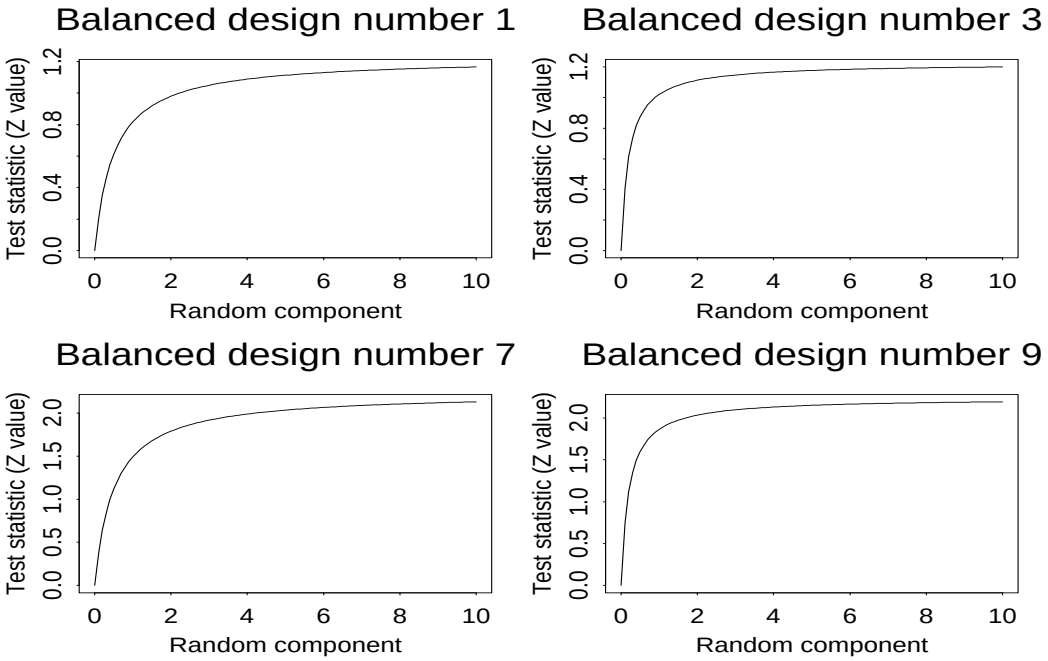


Figure 6.4: Wald test statistics for different σ_a^2 (with $\sigma_e = 0$) under different balanced designs.

$$\frac{1}{2}\,\frac{\sqrt{2}\,z\,(2\,\sqrt{2}\,z + 2\,\sqrt{m})}{(-2\,z^2 + m)\,n}.$$

Using this, we find that each design has an absolute maximum limit for the possible z value. This is particularly visible for designs with a small number of individuals, where this limit becomes quite small (For exsamples of this see figure 6.4). If we extend this limit either by a very small p-value or too few individuals, the differences between σ_a^2 and σ_e^2 will never make a difference, as the Wald test under this scheme never will reject H_0, regardless of the observed data (!).

Another example of the effect of covariance between an estimate and its standard deviation, is shown in figure 6.5. Her we have simulated 1000 datasets with 10 standard normal distributed observations. In this fixed effects case, the estimates and its standard deviations are independent, and as shown in the left figure, the test statistic for positive values is quite close to the right side of the normal distribution. If we now sort these estimates and standard deviations by size, we get the figure on the right side. In this figure, the mean has dropped from 0.90 to 0.74, and we have almost none

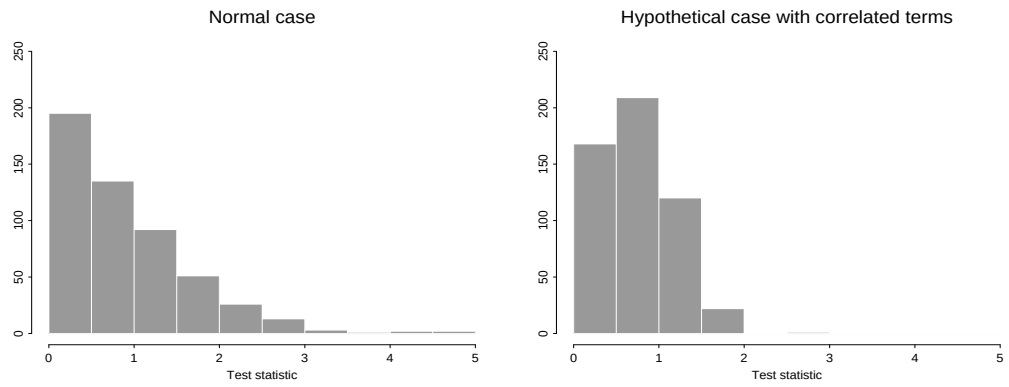


Figure 6.5: Simulated example of test statistics, with and without correlation between the estimates and it's standard deviation. Normal case to left, and the hypothetical case with correlated estimated and it's standard deviation to the left (where the values matched by size). Out of 1000 datasets, figures are just for the 520 positive test statistic, as only these have relevance to one-sided testing.

large values left. This again gives the test very low probability of rejecting H_0.

As we see, the correlation between the estimators and its standard derivation, tells us why the standard Wald test often never rejects H_0 at all for small designs. From the previous section, we also remember the new "Wald - H_0" test. As we in this test set $\sigma_a^2 = 0$, it does not suffer from the same problems of covariance between the estimate and its standard deviations as the standard Wald tests. And as we have seen in the simulations, this method also does not suffer from the same problems of to few rejections of H_0 (but actually have another problem of all too high significant levels).

As an exact demonstration of the problems with correlated standard derivations, we have perform a hypothetical Wald (ML) test without any correlations between the estimate and its standard deviation. This was done by randomly mixing the estimates and its standard deviations from our 5000 simulated datasets. The result is a Wald test with exactly the same distribution of both the estimates and standard deviations, but with no correlations between these values. The practical consequence of this for balanced design number 7, with significant level 0.01, is shown in table 6.9. As we see, the test without correlations often rejects H_0, and even has a somewhat too high significant level.

H_0	.1	.2	1	5	25
0.080	0.110	0.134	0.266	0.412	0.457

Table 6.9: Hypothetical case: 0.01 level Wald (ML) test for balanced design no. 7: {2,2,2,2,2,2,2,2,2,2}, without correlated estimated and standard deviation (E.g. Simulated estimates and standard deviations are randomly matched). In real life this test never rejected H_0.

Why does the Wald test sometimes perform worse with more observations?

Looking at the simulation results in section 5.1.1, we find that the distributions of the variance components are sometimes somewhat skewed, with a tail towards large estimates. Under H_0, this should give higher probability of rejecting the H_0 hypothesis. Still, we often get a very conservative test for small designs, as the correlation between the estimates and its standard deviation (as discussed in the previous section) gives us much smaller probability of rejecting H_0.

In figure 5.3 and 5.4, we see that an increase the number of repetitions gives a distribution of the estimates somewhat closer to the normal distribution. Comparing the different designs shown in figure 6.4, we find that the effect of correlated estimates depends almost totally on the number of individuals, with little impact of the number of repetitions. As we then increase the number of repetitions, we get lower power as the effect of a non normal distribution decreases. Normally, this would give us a more correct test, but since we already are far below the correct significant level, because of correlated estimates and standard deviation, this actual gives us a somewhat less precise, not a better test.

For the simulations, this is shown in a comparison of table 6.4 and 6.6. In practice, this gives a quite special situation, where the Wald test for these designs perform worse with more observations (!). Another example showing that more observations sometimes does not give us a better Wald test, is comparison of table 6.6 and 6.8. Here, the increase in the number of individuals, reduces the effect of correlated estimates and its standard deviations, leaving us with a too high significant level for the larger design (because of the somewhat skewed distribution of the variance component estimate).

Why does the "Wald - H_0" test give so high significant levels?

As seen in section 6.1, the "Wald - H_0" test gives in almost all cases too high level of significance. This could probably mainly be atributed to the way we have estimated σ_e^2. As for the σ_e^2 estimate to be independent of the $\hat{\sigma}_a^2$, it is estimated under the random effects model. The "Wald - H_0" test is only rejected for positive values of $\hat{\sigma}_a^2$, and for these values, the $\hat{\sigma}_a^2$ estimate takes up some of the model variation that is really from σ_e^2. Under H_0, this gives a general underestimation of σ_e^2. As a consequence, we get a too low estimated standard deviation, which again results in too many rejections of H_0.

6.2.2 Why is the likelihood test so conservative?

Looking at table 6.3, we find that the likelihood ratio test performs acceptably, but very conservatively. At first look, this could be blamed on the test's asymptotical nature, but looking at table 6.2 versus table 6.8, we see that the real significance levels are almost unchanged when the number of observations increases (e.g. n and m increases). So why is the asymptotical theory not working properly?

From the likelihood ratio theory, we remember that we should take the maximum log-likelihood under the H_0 limitation, and divide it by the maximum log-likelihood for the full model. But are we really doing this? Have we not introduced the limitation that $\sigma_a^2 > 0$ on the full model?

This can easily be demonstrated by the alternative way of writing the mixed effects formula. From section 1.1, we remember that the mixed effects formula could be written as $N(x\beta, \Sigma)$, where Σ includes both the random effects and the error term. In this model, Σ can be seen as a covariance matrix. This includes the possibility of negative covariances, but in our definition of the random components, we have limited the distribution to positive values.

The effect of allowing negative values in balanced design number 1 to 9, can be seen in table 6.10. As we see, most rejections are really done for negative values. Including these negative values, the real p-values come much closer to the intended p-value. Actually, the likelihood ratio test that allows negative values, is the correct likelihood test in reference to the asymptotical theory, while our test gets a serious boundary problem. A general theory for correction of such problems is presented in Self & Liang (1987), but no such corrections are implemented in today's leading statistical packages.

Design	$p = .05$	$p = .01$
Ballanced no. 1: {2,2,2}	0.127 (0.019)	0.042 (0.003)
Ballanced no. 3: {5,5,5}	0.156 (0.010)	0.056 (0.002)
Ballanced no. 7: {2,2,2,2,2,2,2,2,2,2}	0.071 (0.025)	0.015 (0.004)
Ballanced no. 9: {5,5,5,5,5,5,5,5,5,5}	0.071 (0.016)	0.018 (0.002)

Table 6.10: Proportion of $\sigma_a^2 = 0$ tests rejected under H_0: $\sigma_a = 0$, with rejection for negative values. Results without rejection of negative values are shown in parenthesis for comparison.

For applied statistics, this leads us to assume that most Likelihood Ratio tests for random effects are quite conservative tests. This again shows us how dangerous a choice the Wald test is, as it for the Mississippi river data (see section 6) gives a p-values very much larger than the (probably) somewhat conservative Likelihood ratio test.

Chapter 7

Conclusions

> Science is the systematic classification of experience.
> George Henry Lewes (1817-1878)

7.1 Summary

In this study, we have looked at estimation and testing in the increasingly popular field of mixed effects models. The classical work on evaluating the different estimation methods for these models was done by Swallow & Monaham in 1984. As a start, we have repeated much of this study, and then expanded the scope of the study by including several additional issues. In this work, we have come across some new and quite interesting findings, which have been presented in chapter 5 and 6.

Probably, the most important of these results has come in an area which Swallow & Monaham did not include in their study; Different methods for testing (the existence of random effects) in mixed effects models. In chapter 6, we compared the three most common approaches for such tests; The Wald test, the Likelihood ratio test and the F-test. To my knowledge, no other studies have evaluated the different testing methods used in mixed effects models, so these results should be new to the statistical community.

Not surprisingly, the Likelihood ratio and F-tests perform somewhat better than the Wald-test. Still, the differences shown in this study will probably come as a surprise to many statisticians. In some quite small designs, it is even shown that the Wald test for quite reasonable significant levels never will reject H_0 - regardless of the observed differences. This gets better when we increase the size of the designs, but for moderat designs the Wald test often gives too high significant levels under H_0.

In contrast to earlier belief (Thompson & Beacon 1998, Littell et al. 1996), we show that the Wald test's problems with random effects in small designs is mostly due to correlations between the estimate and its standard deviation estimate, and not to skewed or bounded distributions. This happens as we, in contrast to ordinary fixed effects models, work with variance components that are included in their own standard deviation estimate.

In practice, this makes the Wald test a quite dangerous choice for practical application. Still, the test is implemented in leading statistical packages (such as the SAS system), and is probably in widespread use in many different fields (For an example of this, see Taylor, Pickering, Lord & Pickles 1998). As for the other tests, it is shown that the Likelihood ratio test have some boundary problems for random effects, that lead to a somewhat conservative test.

All these findings are supported both by simulation studies and some theoretical work, and lead to recommendations for applied statisticians and developers of the many new routines for analyzing mixed effects models.

In addition to the comparison of different testing methods, we have also looked at several other issues concerning methods for analyzing mixed effects models. The most important of these results is probably an evaluation of the Bayesian Gibbs sampler for "uninformative priors". In this simulation study, it is shown that the Gibbs sampler converges quite fast to the Maximum Likelihood estimate, as the number of observations increases. For practical purposes, this shows that the Gibbs sampler probably can be a good choice for complex situations where it is difficult to implement standard frequentist methods.

7.2 Recommendations for applied statisticians

In the last chapters, we have studied the performance of many different estimation and testing methods in mixed effects models. To some degree, the differences between these methods have a unique academic interest, but for the applied statistician or University teacher, one question stands out; *Which methods are recommended for practical use in the different situations?*

For developers of statistical software, this question has become especially important. During the last decade, mixed effects models has gone from being a specialized modeling tool only used by professional statisticians, to a quite ordinary method for analyzing data in many different fields and professions. The basic ideas behind mixed effects models are somewhat more advanced

than ordinary regression, but still do not require deep mathematical reasoning. In practice, I believe that the use of this model has been limited rather by the opportunity to perform the analysis, than by the actual understanding of the model. This has now began to change, as leading statistical software packages now are building user friendly graphical interfaces to the basic versions of the mixed effects models. This makes the model available to a much wider audience, but also makes the choice of default methods very important, as many of these new users probably are not familiar with the advantages and limitations of the different methods.

Based on this and the earlier studies, my recommendations to practical statistics and developers of statistical software are presented in the next section.

7.2.1 Recommended estimation methods

In this study, we have seen that even under the correct model, we can experience large differences between the different estimating methods as the designs become unbalanced. In the choice between the different estimation techniques, we must to some degree choose between different priorities:

- Minimizing the mean square error
- Unbiased estimation
- Easy modeling properties
- Fast compilations

Considering these priorities, I want to give the following general advices to users mixed effect models:

- *As a general rule, the Maximum Likelihood (ML) estimator is a good choice.*
- If we want an *unbiased estimate,* we may choose the Restricted Maximum Likelihood (REML) estimator, but it also gives a somewhat larger mean square error than the more traditional Maximum Likelihood (ML) estimator.
- For *fast compilations*, the ANOVA and MINQUE(0) are good choices. On the other hand the MINQE (0) estimator, and to some degree the ANOVA estimators, can sometimes give a considerably larger mean

square error than the Maximum Likelihood (ML) approach. This is most evident for considerably unbalanced models combined with a large ratio of σ_a^2/σ_e^2.

- If we have additional information to the data, the Bayesian Gibbs sampler has the advantage of making it easy to include this information in the model (as priors).

- For many practical applications, the Gibbs sampler will give *especially easy modeling properties*, and even for "noninformative priors" it often gives results quite near the Maximum Likelihood estimate. Still, the estimates rely heavily on the priors for small amounts of data, and the Gibbs sampler is not recommended in the smallest designs without special information available for choosing the priors. These problems can to some degree be reduced by a good choice of "noninformative priors", but the Maximum Likelihood (ML) estimator is overall a much better predictor for small designs.

- Recommendations for developers of statistical software:

  Maximum likelihood (ML) is a good choice as default method, as it performs well in most situations. REML, MINQUE(0) and Gibbs sampler have all its specialities, and could be usefull to implemented as alternatives.

7.2.2 Recommended methods for testing random components

In contrast to different estimation methods, we have a fairly simple goal in evaluating the different testing methods. We want a test with the highest possible power (under H_1), while the true significant level never extends our desired significant level. Based on this, I will give the following advice for applied statistician:

- In contrast to most fixed effects cases, the Wald test will often have very lower power in testing random effects for small designs. On the other side, it sometimes gives too high true significant levels for moderate designs. For this reason, the *Wald test is usually not advisable for testing random effects.* This is particularly true for low significant levels and small samples, where the Wald test sometimes never rejects H_1 at all.

- In the choice between the F and likelihood ratio test, the F test has somewhat better power under H_1. Still, the likelihood ratio test perform fairly well, and it is usually no unnecessary to make an approximated F test in the cases where no F test is available. *In practice, the general advice will be to stick with the Likelihood ratio test, except for the cases where we need a test with extra good power under balanced designs.*

- Recommendations for developers of statistical software:
  As the Likelihood ratio test is available and performs well both in balanced and unbalanced models, I will recommend it as the default testing method. In addition, it may be useful to implement the F-test as an option, as it gives somewhat better power in balanced models.

  As for the practical implementation of the Likelihood ratio test, today's statistical packages mostly implement it by printing the deviance for the full model. With this approach, the user will then have to run the analysis for the reduced models, and then calculate the p-value by hand. With increasingly fast computers, my option is that the time has come for a direct calculation of the Likelihood ratio test. With many parameters, this will require some computing time, but in practice, this will probably be faster for advanced users, and make the test more available for novice users.

7.3 Basis for further studies

Still, several questions remain unanswered and make up good basis for further studies. These questions include:

1. Regarding testing:

   – How well does this simulation study of different testing methods expand to unbalanced designs?
   – In the light of the debate regarding the principles behind testing against zero effect, we might ask how well these simulation results expand to tests with alternative definitions of H_0? (E.g. Tests of clinical relevant differences - No just a difference)
   – Find out whether Monte Carlo tests could be useful for making better testing methods in unbalanced mixed effect models (For an example of a simulation study of Monte Carlo tests, see Dimakos 1995.).

– Take a more detailed look at some theoretical aspects, like the relationship between the likelihood ratio and F tests.

2. Regarding estimation:
   – Further check the assumptions and effect of different priors for the Gibbs sampler. It could also be interesting with a more in depth look at the theoretical sides regarding comparison of frequentist and Bayesian methods.
   – Look at different methods for confidence intervals, such as Bootstrapping and Likelihood based theory.
   – Study estimating (and testing) in a more general setting. This could include more complex linear mixed effect models, multi level models, non-normal distributions, non-normal links and misspecified mixed effects models.
   – Study numerical routines and efficiency for large datasets.

3. See how these results are linked to some practical cases.

Appendix A

Computer algorithms and "tricks"

As in all simulation studies, this thesis is based on a considerable amount of programming. In this appendix, we will shed some light on this part of the work, and reveal some "tricks from behind the scenes".

A.1 General programming principles

As seen in table A.1, the basic simulations alone are done for 11 different estimating and testing methods. These methods are combined with several designs, for a total of 95 combinations. Under each of these combinations, we again have different values of σ_a^2/σ_e^2. In addition to this, we still have some methods such as the likelihood test without the $\sigma_a^2 > 0$ restriction, that are not included in these numbers. All in all, this makes up several hundred cases under which we have simulated just for these published data.

In the process, we have also performed simulations on several other combinations and models. Examples of this is a study of mixed effects model with random regression parameter, and a study of the effect on σ_e^2 estimates

	Number of methods	Designs	Combinations of σ_a^2/σ_e^2
Frequentist estimation	4	13	4
Testing	5	7	6
Bayesian estimation	2	4	4

Table A.1: Basic simulations for this study.

from truncation of negative values of σ_a^2. In the end, these studies did not give results which were interesting enough to be included in this thesis.

As we see, there has been a large number of different situations to simulate, and in this process there are especially three important issues:

1. How can we minimize the probability of programming errors?
2. How can we reduce the need for computer power (computing time)?
3. How can we ease the work of performing the simulations?

The first issue can be dealt with using different tests of consistency with known theory and earlier studies. In this thesis several such tests are performed. Some examples are:

- The procedures for simulating random effects data, are tested by drawing large number of observations and comparing row and column differences with their theoretical expectations.
- The procedures for frequentist estimates are compared to results in Swallow & Monaham (1984).
- Several procedures are tested for large scale data and compared to limit theorems.

Another method of minimizing the probability of programming errors, are through the use of pre-programmed and well tested packages. Many of our calculations have been done using such procedures.

In addition to these tests, we *have focused strongly on making functions and objects modular and reusable*. Using these principles, the source code gets much smaller and more perspicuous. This eases the work tremendously, and gives tests mentioned earlier a much wider validity. To implement these principles, we have largely used the object-oriented S language for both the simulations and the following analysis. This language is today incorporated in the S-PLUS and R statistical packages, and the inventor John M. Chambers expresses the goal of the S language as "To turn ideas into software, quickly and faithfully". For more information about the S language, see Venables & Ripley (1997) and Chambers (1998).

In order to reduce the need for computer power, we used binary routines for all computer intensive tasks like the REML, ML and Gibbs sampler estimators. This is done as the S interpreters are quite slow compared to binary

code. For all these tasks, we have found pre-coded routines, and have chosen to use them as they probably are quite well optimized.

In my S routines, there are also to some degree a choice between making fast specialized functions, or making general reusable functions. In this choice, we have mostly preferred making the routines reusable. This is done since it in this study only has a small impact on the total consumption of computing time, as most of the time is used in the computer intensive binary routines. For the ML and REML estimating routines, we could have used sufficience principles as shown by Swallow & Monham (1984), but in this study we have found this not to be worth the extra work and the problems with a less surveyable and well tested source code.

A.2 Simulating datasets with random effects

Using the principles of modular and reusable source code, all datasets (balanced and unbalanced) are generated using the same S function. This function is shown below:

```
''mixnub'' <- function(x, antm, alfav, s)  {
  # X      - The fixed effect.
  # Antm   - A vector of reputations for the different
  #          individuals. To simulated serval datasets,
  #          repeat this vector with ``rep(c(...),n)''.
  # alfav  - The standard deviation of the random effect.
  # s      - The standard deviation of the error term.
  maxantm <- max(antm)
  n <- length(antm)
  alfabeta <- rep(1,maxantm)
  res <- t(matrix(rep(x * alfabeta,n),maxantm))
  # ``res'' is the matrix for the simulation results.

  # We simulated the random effect:
  if ((alfav==0)==F) {
    z <- matrix(rnorm(n,0,1),n)
    res <- res + z %*% alfabeta * alfav
  }

  # We simulated the error term:
  if (s>0) {
```

```
    for(i in 1:n) {
        res[i,1:antm[i]]<-res[i,1:antm[i]]+rnorm(antm[i],0,s)
        if (antm[i]<maxantm)
          res[i,(antm[i]+1):maxantm] <- NA
    }
  }
  else {
    for(i in 1:n)
      if (antm[i]<maxantm) res[i,(antm[i]+1):maxantm] <- NA
  }

  return(res)
}
```

To generate two sets of data from the unbalanced design number 1 ($n_i = \{3, 5, 7\}$) with $\mu = 3$, $\sigma_a = 0.1$ and $\sigma_e = 1$, we can now use the following command:

```
> mixnub(3,rep(c(3,5,7),2),.1,1)
          [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]
[1,] 1.753024 3.845164 2.106750       NA       NA       NA       NA
[2,] 2.621491 1.263815 3.599750 1.523219 2.618191       NA       NA
[3,] 2.275264 2.034738 4.006103 2.140304 2.574507 2.917348 2.883158
[4,] 2.099634 3.081872 2.248927       NA       NA       NA       NA
[5,] 3.370456 2.076941 1.849028 2.715241 3.415091       NA       NA
[6,] 2.909810 1.145729 2.519810 2.171350 3.443843 3.526429 3.462820
```

These data are good for matrix operations, but several pre-programmed S functions need data as S dataframes. The following function is used for this convention:

```
mix <- function(mixres) {
  n <- dim(mixres)[1]
  m <- dim(mixres)[2]
  res <- matrix(c(rep(0, times = (n * m * 2))), n * m,)
  res[, 1] <- sort(rep((1:n), times = m))
  res[, 2] <- as.vector(t(mixres))
  res <- data.frame(res)
  names(res) <- c(''subj'', ''ysim'')
```

Estimate	Calculation metode
Balanced case estimates	Custom built matrix operations
ML & REML	varcomp function & lme-rutine (Pinheiro & Bates 1995)
MINQUE(0)	varcomp function
ANOVA	Custom built matrix operations
Wald-tests	Custom built matrix operations
F-test	Custom built matrix operations
LR-tests	lme-routine (Pinheiro & Bates 1995) and some custom built matrix operations

Table A.2: S-PLUS routines used in this study.

```
  return(res)
}
```

A.3 Functions for frequentist estimation and testing

All frequentist estimation and testing is done using the S-PLUS package, which is a super set of the S language. How the different estimates are calculated are shown in figure A.2. As we see, several functions are custom-built, and all the functions take up around 450 lines of S code. To both simplify the code and reduce the calculations, the Wald-tests uses results from the estimation. Still, the implementation of testing methods is quite complex, and takes up around 320 of the 450 lines of code.

The S language is usually quite slow for simulations, as both implementations available today (S-PLUS and R) only include S as an interpreted language. For this reason, most simulation studies using S, includes their

own user generated binary code for numerical intense tasks. This binary code is made by programming languages such as C or C++, and is called from within the S language. This is generally the advisable approach, but in this study all the numerically intensive tastes like ML, REML etc. could be found in the S-PLUS package. These functions are probably well optimized, so there was no need for a custom build binary code. ANOVA, Wald-test etc. are all programed via matrix operations, that are calculated quite rapidly using the S language.

As an example of a custom built S function, "finnall3" is shown below. This function is the basic function for estimation in the unbalanced case.

```
''findall3'' <- function(mixdata,design) {
  # mixdata   - Data generated using the mixnub function.
  # design    - Specification of the design as
  #             a vector of the number of repetitions.
  obs      <- sum(design)
  m        <- length(design)
  obs2     <- m*max(design)
  ant      <- dim(mixdata)[1]/m
  frig     <- obs - m
  konsd    <- sum(design^2)
  frig2    <- obs - konsd/obs
  res      <- matrix(0, ant,8)
  mixdataf <- mix(mixdata)
  datamix  <- data.frame(subj=factor(mixdataf$subj)
              ,ysim=mixdataf$ysim)
  is.random(datamix$subj) <- T

  for(i in 1:ant) {
    datamix2 <- datamix[(1+(i-1)*obs2):(i*obs2),]
    datamix2 <- datamix2[(datamix2[,2]!=''NA''),]
    # Maximum Likelihood estimate:
    resml <- varcomp(ysim~subj,datamix2,method=''ml'')
    res[i,1] <- resml$variances[1]
    res[i,2] <- resml$variances[2]
    # Restricted Maximum Likelihood estimate:
    resreml <- varcomp(ysim~subj,datamix2,method=''reml'')
    res[i,3] <- resreml$variances[1]
    res[i,4] <- resreml$variances[2]
```

```
    # MINQUE(0) estimate:
    resminq <- varcomp(ysim~subj,datamix2)
    res[i,5] <- max(0,resminq$variances[1])
    res[i,6] <- max(0,resminq$variances[2])
    # ANOVA estimate:
    mixd <- mixdata[(1+(i-1)*m):(i*m),]
    res[i,8] <- sum((mixd - apply(mixd,1,mean,na.rm=T))^2
                ,na.rm=T)/frig
    res[i,7] <- max(0,(sum((apply(mixd,1,mean,na.rm=T)
                -mean(mixd,na.rm=T))^2
                *design)-(m-1)*res[i,8])/frig2)
  }
  return(res)
}
```

A.4 Bayesian estimating - Using the BUGS package

Today, S-PLUS has no functions for Bayesian estimation, and implementing the Gibbs sampler in S would have resulted in a huge consumption of computer time. An alternative could have been a general programming language such as C, but to get a fast and stable implementation of the Gibbs sampler, this would have required a considerable amount of programming.

As an alternative, we have the "classical BUGS" software package for UNIX. This is a very fast implementation of the Gibbs sampler, but it is made for estimating one dataset at a time. To come around this problem, we have used a little trick; *All the simulated data for one combination (of design and parameters) are imported as one dataset, and the model is specified as identical independent parts that makes up one model for each dataset.* In this way, we get the total simulation problem on a multidimensional level, which can be solved as one large model by the BUGS package.

As an example of this, the BUGS routines for the "default" priors of unbalanced design number 1 ($n_i = \{3, 5, 7\}$) are shown below.

```
model smixef;
```

```
const
  BATCHES = 3,  # Number of individuals (n).
  SAMPLES = 7,  # Maximum number of observations (m).
  NRSIM = 5000, # Number of simulations.
  OBS = 15000;  # Total number of batches (= BATCHES*NRSIM).

var
  y[OBS,SAMPLES], mu[BATCHES,NRSIM], b[BATCHES,NRSIM],
  theta[NRSIM], tau.within[NRSIM], tau.between[NRSIM],
  sigma2.within[NRSIM], sigma2.between[NRSIM];

data y in ''simdata.txt'';
inits in ''smixef.in'';

# Model spesification:
for (m in 1:NRSIM) {
  for (i in 1:BATCHES) {
    for (j in 1:SAMPLES) {
      y[i+((m-1)*BATCHES),j] ~dnorm(b[i,m], tau.within[m]);
    }
    b[i,m] <- theta[m] + mu[i,m];
    mu[i,m] ~dnorm(0, tau.between[m]);
  }
}

# Priors:
for (n in 1:NRSIM) {
  theta[n] ~dnorm(0.0, 1.0E-10);
  tau.within[n] ~dgamma(0.001, 0.001);
  sigma2.within[n] <- 1/tau.within[n];
  tau.between[n] ~dgamma(0.001, 0.001);
  sigma2.between[n] <- 1/tau.between[n];
}
```

To use this model specification, we export the simulated S-PLUS datasets to plain text files and supply one file with priors. This code is then run in BUGS using the following BUGS commands:

```
compile(''smixef.bug'')
update(1000)
```

```
monitor(sigma2.between,10)
update(10000)
stats(sigma2.between)
q()
```

At last, the results are imported in S-PLUS for further analysis.

A.5 Running the simulations

The functions mentioned earlier give us methods for performing the simulations, but running actual simulations still requires a large list of commands. As an example, only the repetition of parts of Swallow & Monaham 1984 (see section 5.1), needed 764 lines of S code (scripts). (Remark: This is partly because the simulation had to be divided into different S calls, to solve S-PLUS memory management problems).

As we can see, this is a considerable amount of work and raises problems of possible misprints. As a solution to these problems, several of the S scripts have been generated via C-programs. This eases the work, and minimizes the probablity of mistyping. All these programs are around 100 lines of code.

A.6 Analyzing the results

As in the case of the frequentist estimation, the analyzing is done using S-PLUS with widespread use of custom made functions. These functions include several hundred lines of code, and one example is shown below:

```
''plotbal''<-function(simdata,m,n,sizeofrc) {
# This function plots the distributions of the ML estimator
# for balanced data. The function plots distributions for 4
# different random components under one design, and the results
# are compared to the corresponding normal distribution.
#
# simdata     List of vectors of REML estimates.
# m           Number of repetitions.
# n           Number of observations.
# sizeofrc    List of sizes for the true random component.

# Divides the screen into 2x2 blocks:
```

```
  screens <- split.screen(figs = c(2,2))
  for(i in 1:4) {
    screen(screens[i])
    simd <- remlml(simdata[[i]][,3:4],m,n)
    # ``remlml'' is a custom built function to convert
    # estimates from REML to ML for the balanced design.
    #
    # We use the S-PLUS function ``density'' for
    # estimation the density of the simulated estimates:
    d1 <- density(simd)
    # We find the corresponding normal distribution:
    dn <- dnorm(d1$x,mean(simd),sqrt(var(simd)))
    # We use ``matplot'' to plot both graphs:
    matplot(matrix(rep(d1$x,2),50),matrix(c(d1$y,dn),50)
            ,type=''l'',xlab=paste(''Random compoent = '',
             as.character(sizeofrc[i])),ylab=''Density'')
  }
  close.screen(, T)
}
```

Appendix B

Computer resources

B.1 Computer systems used in the study

Most of the simulations have been done using S-PLUS under the IBM AIX platform (UNIX). One notable exception is the Gibbs sampling, which has been done using the BUGS software (Spiegelhalter, Thomas, Best & Gilks 1995) on both Linux and AIX platforms. Plots and figures are created in S-PLUS under HP-UX UNIX, Linux and Windows NT.

B.2 Software & dataset references

Some statistical programs with mixed effects models

- BUGS:
  http://www.mrc-bsu.cam.ac.uk/bugs/
- R:
  http://stat.auckland.ac.nz/r/r.html
- SAS:
  http://www.sas.com/
- S-PLUS:
  http://www.mathsoft.com/splus/
- SPSS:
  http://www.spss.com/
- Statistica:
  http://www.statsoft.com/

Mixed effects add on routines:

- NLME / LME for R:
  http://www.ci.tuwien.ac.at/R/src/contrib/PACKAGES.html
- NLME / LME for S-PLUS:
  http://cm.bell-labs.com/cm/ms/departments/sia/project/nlme/

General longitudinal data routine for S-PLUS:
(including mixed effects)

- OSWALD for S-plus
  http://www.maths.lancs.ac.uk/Software/Oswald/

Datasets used in this study

- Datasets are from "SAS System for Mixed Models" (Littell et al. 1996), and can be found at:
  ftp://ftp.sas.com/pub/publications/A55235

B.3 Acknowledgments

From The Research Council of Norway (Program for Supercomputing), this work has received a grant of computing time on the supercomputing facilities at Oslo University. This grant of easily accessible computer power, has made my work considerably easier. I wish to thank The Research Council of Norway for its support of the project. I also want to thank Section of Medical Statistics for making a superb office with a Windows NT workstation (with X-vision) available to me while I worked on this thesis.

List of Figures

List of Tables

Bibliography

Chaan, A., Laird, N. M. & Slasor, P. (1997), 'Tutoral in biostatistics; using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data', *Statistics in medicine* **16**, 2349–2380.

Chambers, J. M. (1998), *Programming with Data. A Guide to the S Language*, Springer.

Data Analysis Products Division MathSoft Inc. (1997), *S-PLUS 4 Guide to Statistics*, MathSoft Inc.

Dimakos, X. (1995), Contributions to monte carlo testing, Master's thesis, University of Oslo.

Engel, B. & Buist, W. (1996), 'Analysis of a generalized linear mixed model: A case study and simulation results', *Biometrical Journal. Journal of Mathematical Methods in Biosciences* **38**, 61–80.

Giesbrecht, F. G. & Burns, J. C. (1985), 'Two-stage analysis based on a mixed model: Large-sample asymptotic theory and small-sample simulation results', *Biometrics* **41**, 477–486.

Gilks, W., Clayton, D., Spiegelhalter, D., Best, N., Sharples, L. & Kirby, A. (1993), 'Modelling complexity: Application of gibbs sampling in medicine', *Journal of the Royal Statistical Sociey* **B 55**, 39–102.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (1996), *Markov Chain Monte Carlo in Practice*, Chapman & Hall.

Hartley, H. O. & Rao, J. N. K. (1967), 'Maximum-likelihood estimation for the mixed analysis of variance model', *Biometrika* **54**, 93– 108.

Hartley, H. O., Rao, J. N. K. & LaMotte, L. R. (1978), 'A simple 'synthesis'-based method of variance component estimation', *Biometrics* **34**, 233–242.

Harville, D. A. (1977), 'Maximum likelihood approaches to variance component estimation and to related problems', *Journal of the American Statistical Association* **72**, 320 – 340.

Henderson, C. R. (1953), 'Estimation of variance and covariance components', *Biometrics* **15**, 192–218.

Hocking, R. R. (1996), *Methods and Applications of Linear Models*, John Wiley & Sons, inc.

Jiang, J. (1996), 'REML estimation: Asymptotic behavior and related topics', *Annals of Statistics* **24**, 255– 286.

Khuri, A. I., Mathew, T. & Sinha, B. K. (1998), *Statistical Tests for Mixed Linear Models*, John Wiley & Sons, Inc.

Lindstrom, M. J. & Bates, D. M. (1988), 'Newton-raphson and EM algorithms for linear mixed-effects models for repeated measures', *Journal of the American Statistical Association* **83**, 1014–1022.

Littell, R. C., Milliken, G. A., Stroup, W. W. & Wolfinger, R. D. (1996), *SAS System for Mixed Models*, SAS Institute Inc.

Maddala, G. S. & Mount, T. D. (1973), 'A comparative study of alternative estimators for variance components models used in econometric applications', *Journal of the American Statistical Association* **68**, 324–328.

Miller, J. J. (1979), 'Maximum likelihood estimation of variance components: A monte carlo study', *Journal of Statistical Computing and Simulation* **8**, 175–190.

Patterson, H. & Thompson, R. (1971), 'Recovery of inter-block information when block sizes are unequal', *Biometrika* **58**, 545 – 554.

Pinheiro, J. C. & Bates, D. M. (1995), Lme and nlme, Technical report, University of Wisconsin-Madison.

Pinheiro, J. C. & Bates, D. M. (1997), Future directions in software for mixed-effects models: Version 3.0 of NLME, *in* 'International S-PLUS User Conference'.

Rao, C. R. (1971*a*), 'Estimation of variance and covariance components; MINQUE theory', *Journal of Multivariate Analysis* **1**, 257–275.

Rao, C. R. (1971*b*), 'Minimum variance and quadratic unbiased estimation of variance componenets', *Journal of Multivariate Analysis* **1**, 445–456.

Rønnevik, P. R. (1999), Statistisk meta-analyse av lungekreft og fett, Master's thesis, University of Oslo, Norway. (Meta analyses of breast cancer and fat).

Searle, S. R. (1970), 'Large sample variances of maximum likelihood estimators of variance componentes using unbalanced data', *Biometrics* pp. 505–524.

Searle, S. R. (1971), *Linear Models*, John Wiley & Sons.

Searle, S. R. (1987), *Linear Models for Unbalanced Data*, John Wiley & Sons.

Searle, S. R. & Henderson, H. V. (1979), 'Dispersion matrices for variance componets models', *Journal American Statistical Association* **74**, 465–470.

Self, S. G. & Liang, K.-Y. (1987), 'Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions', *Journal of the Ameriacan Statistical Association* pp. 605– 610.

Spiegelhalter, D. J., Thomas, A., Best, N. G. & Gilks, W. R. (1995), Bayesian inference using gibbs sampling, version 0.50, Technical report, Cambrigde MCR Biostatistics Unit.

Spiegelhalter, D. J., Thomas, A., G., B. N. & Gilks, W. R. (1996), Bugs 0.5 examples - volume 1 (version i), Technical report, Cambrigde MCR Biostatistics Unit.

Swallow, W. H. (1981), 'Variances of locally minimum variance quadratic unbiased estimators ("MINQUE's") of variance components', *Technometrics* **23**, 271–283.

Swallow, W. H. & Monahan, J. F. (1984), 'Monte carlo comparison of ANOVA, MINQUE, REML, and ML estimators of variance components', *Technometrics* **26**, 47–57.

Swallow, W. H. & Searle, S. R. (1978), 'Minimum variance quadratic unbiased estimaton (MINQUE) of variance components', *Technometrics* **20**, 265–272.

Taylor, A., Pickering, K., Lord, C. & Pickles, A. (1998), *Statistical Analysis of Medical Data - New Developments*, Arnold, chapter Mixed and multilevel models for longitudinal data: growth curve models of language development, pp. 127 – 143. Brian S. Evritt (eds.) and Graham Dunn (eds.).

Thompson, S. G. & Beacon, H. J. (1998), *Encyclopedia of Biostatistics*, Wiley, chapter Mixed Effects Models for Longitudinal Data. Peter Armitage (eds.) and Theodore Colton (eds.).

Venables, W. & Ripley, B. (1997), *Modern Applied Statistics with S-PLUS. Second Edition*, Springer.

Yu, H., Searle, S. R. & McCulloch, C. E. (1994), 'Properties of maximum likelihood estimators of variance components in the one-way classification, balanced data', *Communications in Statistics, Part B - Simulation and Computaing* **23**, 897– 914.

About this thesis:

In this thesis, we have studied statistical methods for analyzing data with repeated measurements. Such data occur in a large number of fields from medicine to economics and technology. An example is the clinical AIDS trial mentioned in section 1.1, where the development of the disease is measured over time for each patient.

With repeated measurements, we often get data that do not fulfill the independent requirements of most standard statistical methods. For the study of these kinds of data, the most correct model is in many cases the "Mixed Effects Model". For this model, we have many different methods for estimation and testing. In this thesis, we perform a study where we compare some of these methods through simulations. In several cases, this study reveals large differences between the different methods. This findings are also supported by some new theoretical considerations.

An example is the comparison of the Wald and the Likelihood ratio tests found in chapter 6. Here, the default method in the leading statistical package SAS, needs several times as many data to reveal true differences as the alternative Likelihood ratio test. Still, this thesis shows that even the Likelihood ratio test is, in practice, a quite conservative test.

Other interesting results includes a study of the increasingly popular Gibbs sampler. In this study, we show that the Gibbs sampler with "uninformative priors" quite fast converges to the Maximum Likelihood estimate, as the number of observations increases. Based on these and other results, I conclude with advises for applied statisticians, university teachers, researchers from other fields, producers of statistical packages and others who are using mixed effects models in their daily work.

Harald Fekjær,
Section of medical statistics, University of Oslo, June 1999

How you gather, manage, and use information will determine whether you win or lose.

"Know your numbers" is a fundamental precept of business.

(From "Business @ the Speed of Thought", Warner Books, 1999.)
William (Bill) H. Gates (1955-)