QG Notes

Peter Sørensen

2022-02-03

Contents

1 Prerequisites **5**

2 Introduction **7**

3 Quantitative Genetics **9**

3.1 Infinitesimal model 10

4 Estimation of Genetic Parameters **19**

4.1 Genetic model 20

4.2 Genetic parameters 20

4.3 Data required for estimating genetic parameters 21

4.4 Statistical models and variance components 22

5 Methods for estimation of genetic parameters **25**

5.1 Estimating heritability using parent - offspring regression 25

5.2 Estimating heritability using ANOVA for family data 27

5.3 Estimating heritability using Restricted Maximum Likelihood . . 31

5.4 When to estimate variance components? 35

6 Applications **37**

6.1 Example one 37

6.2 Example two 37

7 Final Words **39**

Chapter 1

Prerequisites

This is a *sample* book written in **Markdown**. You can use anything that Pandoc's Markdown supports, e.g., a math equation $a^2 + b^2 = c^2$.

The **bookdown** package can be installed from CRAN or Github:

```
install.packages("bookdown")
# or the development version
# devtools::install_github("rstudio/bookdown")
```

Remember each Rmd file contains one and only one chapter, and a chapter is defined by the first-level heading `#`.

To compile this example to PDF, you need XeLaTeX. You are recommended to install TinyTeX (which includes XeLaTeX): https://yihui.org/tinytex/.

Chapter 2

Introduction

You can label chapter and section titles using `{#label}` after them, e.g., we can reference Chapter 2. If you do not manually label them, there will be automatic labels anyway, e.g., Chapter **??**.

Figures and tables with captions will be placed in `figure` and `table` environments, respectively.

```
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```

Reference a figure by its code chunk label with the `fig:` prefix, e.g., see Figure 2.1. Similarly, you can reference tables generated from `knitr::kable()`, e.g., see Table 2.1.

```
knitr::kable(
  head(iris, 20), caption = 'Here is a nice table!',
  booktabs = TRUE
)
```

You can write citations, too. For example, we are using the **bookdown** package (Xie, 2020) in this sample book, which was built on top of R Markdown and **knitr** (Xie, 2015).

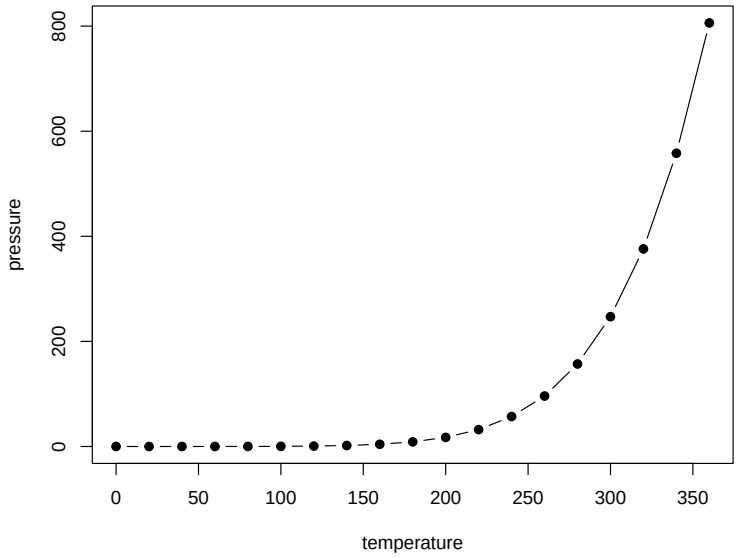

Figure 2.1: Here is a nice figure!

Table 2.1: Here is a nice table!

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa

Chapter 3

Quantitative Genetics

This section introduces the basic concepts in quantitative genetics such as:

- Genetic value and variance for a quantitative trait
- Genetic parameters (heritability, genetic variance and correlation)
- Difference between genetic values and breeding values
- Infinitesimal model

These concepts are relevant for a range of genetic and statistical analyses of complex traits and diseases in animal and plant populations, including:

- Estimating the effect of single locus (or marker) for gene discovery
- Estimating the effect of multiple loci (or markers) for genomic prediction
- Estimating the heritability of a trait (the part of its variability due to genetics)
- Estimating breeding values by pedigree or genomic information
- Selection of breeding individuals based on estimated breeding values
- Prediction of selection response based on estimated heritability (breeder's equation)

In the appendix, further details about the quantative genetic models are presented, but these may be outside the scope of this BSc course.

Quantitative genetics, also referred to as the genetics of complex traits, is the study of quantitative traits. Quantitative genetics is based on models in which many genes influence the trait, and in which non-genetic factors may also be important. Quantitative traits such as size, obesity or longevity vary greatly among individuals. Their phenotypes are continuously distributed phenotypes and do not show simple Mendelian inheritance (i.e., their phenotypes are distributed in discrete categories determined by one or a few genes). The quantitative genetics framework can also be used to analyze discrete traits like litter

size (which consist of discrete counts like 0, 1, 2, 3, ...) or binary traits like survival to adulthood (which consist of 0 or 1, 'dead' or 'alive', etc.), provided that they have a polygenic basis (i.e., they are determined by many genes). The quantitative genetics approach has diverse applications: it is fundamental to an understanding of variation and covariation among relatives in natural and managed populations; it is also used as basis for selective breeding methods in animal and plant populations (https://doi.org/10.1098/rstb.2009.0203).

3.1 Infinitesimal model

The infinitesimal model, also known as the polygenic model, is a widely used genetic model in quantitative genetics. Originally developed in 1918 by Ronald Fisher, it is based on the idea that variation in a quantitative trait is influenced by an infinitely large number of genes, each of which makes an infinitely small (infinitesimal) contribution to the phenotype, as well as by environmental (non-genetic) factors. In the most basic model the phenotype (P) is the sum of genetic effects (G), and environmental effect (E):

$$P = G + E \tag{3.1}$$

The genetic effect (G) in the model can be split into additive effects (A), dominance effects (D), and epistatic effects (I) such that the expanded infinitesimal model becomes:

$$P = A + D + I + E \tag{3.2}$$

The genetic effect may also depend on the environment in which they are expressed (e.g., in plants a drought-tolerance gene may have a favorable effect on grain yield under water-limited conditions, but may be useless under irrigation). Therefore we may consider an extended version of the infinitesimal model where the phenotype (P) is the sum of genetic effects (G), environmental effect (E), and genotype-environment interaction effects (G×E):

$$P = G + E + GxE \tag{3.3}$$

In practice, the genotype-environment interaction effect can be important for the phenotype of individuals, but for the sake of simplicity we will ignore them in the remainder of this section. Therefore, hereafter, we will assume that genotypic effects are not impacted by environmental factors.

3.1.1 Genetic effects, genetic values and breeding values

The genetic effect (G) in the model can include additive effects (A), dominance effects (D), and epistatic effects (I). Additive effects are the summed effects of individual alleles. Dominance effects are interactions between alleles within loci. Epistatic effect are interactions between alleles in different loci and can therefore only occur if two or more loci affect the trait.

Consider an individual that is diploid, like most animals and plants like maize, soybean, barley (i.e., they carry two copies of every genes, except in their sexual chromosomes). Assume that one locus in its genome exists under two possible alleles: A1 and A2, with respective allele effects +1 and -1. How do the individual's alleles combine into a genotype? They may combine additively, so that the value of a genotype (the combination of two alleles genotype) is simply the sum of allele effects, but this is only a very special case! If genetic effects are entirely additive, then the value of each possible genotype is the sum of their respective allele effects, i.e., -2 if the individual is A2A2, 0 if it is A2A1 (or A1A2), and +2 if it is A1A1. Generally, the value of each genotype will depend on the combination of alleles within one locus (G = A + D) or across multiple loci (G = A + D + I). For example, in presence of dominance, the value of each possible genotype may be -2 if the individual is A2A2, +1 if it is A2A1 (or A1A2), and +2 if it is A1A1.

3.1.1.1 Additive Effects

Additive effects are the summed effects of average allele effects. Quite confusingly, additive effects depend on the population, because average allele effects depend on the frequency of genotypes in the population! For example, assume that genotypes have values -2 (A2A2), +1 (A2A1) and +2 (A1A1). In a population consisting of 25% A2A2, 50% A2A1 and 25% A1A1, you would expect the A1 allele in a A1A2 genotype 2/3 of the time, and you would expect the A1 allele in a A1A1 genotype 1/3 of the time. In another population consisting of 90% A2A2, 18% A2A1 and 1% A1A1, you would expect the A1 allele in a A1A2 genotype about 95% of the time, and in a A1A1 genotype only about 5% of the time. As a result, the effect of the A1 allele, averaged over genotypes, will not be the same, from one population to another. The concept of additive genetic effects and average allele effects is fundamental to quantitative genetics. However, it is one of is most confusing, precisely because of the dependance of allele effects on genotype frequencies.

3.1.1.2 Dominance Effects

Dominance genetic effects are the interactions among alleles at a given locus. This is an effect that is extra to the sum of the additive allele effects. Each genotype has its own dominance effect, denoted by δ_{ij}, for the specific combination

of alleles i and j, (e.g., $\delta_{A_1A_2}$), and each of them are non-zero quantities. Using the previous example, the additive and dominance effects would give …..to be completed

3.1.1.3 Epistatic Genetic Effects

Epistatic genetic effects encompass all possible interactions among the loci impacting the trait, whenever there is more than one such loci. This includes all two-way interactions (e.g., interactions between loci A and B, A and C), three-way interactions (e.g., joint interaction among A, B and C), etc. Epistasis can be decomposed, so it includes interactions between additive effects at different loci, interactions between additive effects at one locus with dominance effects at a second locus, and interactions between dominance effects at different loci.

3.1.1.4 Genetic value versus Breeding value

For selective breeding purposes additive genetic effects are of primary interest. This is because additive effects are generally the largest of the genetic effects, and the allelic effects are passed directly to offspring while the other genetic effects are not transmitted to the progeny, and are generally smaller in magnitude. The sum of the additive effects of all loci on a quantitative trait is known as the true breeding value.

- Breeding value = the value of genes to progeny (additive effects only)
- Genetic value = the value of genes to self (which includes additive, dominance and epistatic effects)

The difference between genetic value and breeding value is largely largely dominance deviation. This is because an individual can express dominance deviation (e.g. an A1A2 heterozygote). However, an individual cannot pass on dominance deviation to its progeny as it only transmits one allele (e.g., an A1A2 heterozygote will either transmit a A1 gamete or an A2 gamete to one of its progeny, but not both!) With fully inbred lines, offspring have the same genotype as their parent, and hence the entire parental genotypic value G is passed along. Hence, favorable interactions between alleles are not lost by randomization under random mating but rather passed along. When offspring are generated by crossing (or random mating), each parent contributes a single allele at each locus to its offspring, and hence only passes along a part of its genotypic value. This part is determined by the average effect of the allele. However, any favorable interaction between alleles is not passed along to their offspring.

3.1.2 Infinite number of loci each with small effect on the phenotype

Quantitative traits do not behave according to simple Mendelian inheritance laws. More specifically, their inheritance cannot be explained by the genetic segregation of one or a few genes. Even though Mendelian inheritance laws accurately depict the segregation of genotypes in a population, they are not tractable with the large number of genes which typically affect quantitative traits. To better understand the infinitesimal model assume Mendelian inheritance to occur at every locus in the genome. Let's say there are 30,000 gene loci in the genome. The number of alleles at each locus varies from 2 to 30 or more. If we assume that there are only two alleles (3 possible genotypes) per locus, and gene loci segregate independently, then the number of possible genotypes (considering all loci simultaneously) would be 3^{30000} which is large enough to give the illusion of an infinite number of loci. Furthermore each of these loci could contribute additive and dominance effects in addition to interaction effects.

3.1.2.1 Distribution of genetic and phenotypic values in single locus model

First we will consider how to model the genetic basis of a quantitative trait when a single locus affects the trait of interest. We call this a single-locus model. The distribution of the genetic values for a set of individuals will be discrete. The frequency of the genetic values depend on genotype frequencies, which in turn depend on allele frequencies of A_1 and A_2. The phenotype is however also influenced by the environment. If we assume that the environmental effects are normally distributed (e.g. $\mathcal{N}(0, \sigma^2 = 1)$) then we can observe that the phenotype distribution is infact normally distributed.

3.1.2.2 Distribution of genetic and phenotypic values in multiple loci model

Now we will consider a multiple-locus model. When several loci are causal (i.e., they have an effect on a certain trait), then we talk about a **polygenic model**. Letting the number of causal loci tend to infinity, the resulting model is called an **infinitesimal model**. From a statistical point of view, the breeding values in an infinitesimal model are considered random with a known distribution. Due to the central limit theorem, this distribution tends to a normal distribution, because of the infinitely large number of causal loci. The central limit theorem says that the distribution of any sum of a large number of very small effects converges to a normal distribution. In our case where a given trait of interest is thought to be influenced by a large number of genetic loci each having a small effect, the sum of the breeding values of all loci together can be approximated by a normal distribution. The histograms below show a better approximation to the

normal distribution for breeding values (summed allele effects at causal loci), as the number of causal loci increases. In practice, 100 independently segregating causal loci may be large enough, so that the infinitesimal model (and the normal approximation is genomic models) is accurate enough for predictions.

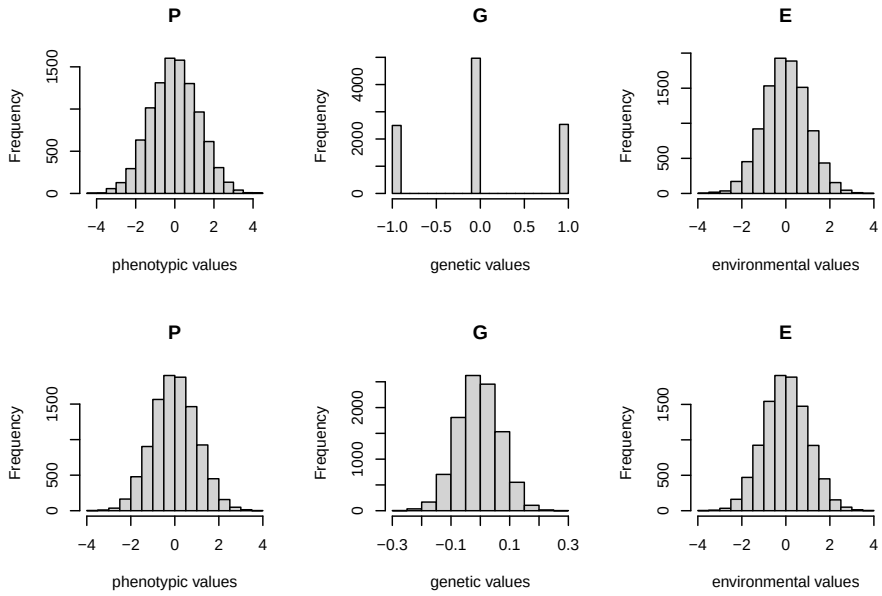

Figure 3.1: Distribution of genetic and phenotypic values for a quantitative trait influenced by a single locus model (top panel) or multiple loci (bottom panel)

3.1.3 Genetic parameters

Fisher (1918) and Wright (1921) have introduced fundamental statistical methods in quantitative genetics:

- analysis of variance: the partition of phenotypic variation into heritable (A) and non-heritable components (D, I and E).
- resemblance among relatives: the estimations of the proportion of loci shared by relatives under the infinitesimal model.

3.1.3.1 Genetic variance:

In the model proposed by Fisher (1918), Cockerham (1954) and Kempthorne (1954), covariance among relatives is described in terms of the additive genetic variance V_A (variance of additive genetic effects, or breeding values), dominance

variance V_D (variance of interaction effects between alleles in the same locus), and epistatic variance V_{AA}, V_{AD}, V_{DD}, …. (variance of interaction effects – additive and/or dominance effects – among loci) (Falconer & Mackay 1996; Lynch & Walsh 1998). These partitions are not dependent on numbers of genes or how they interact, but in practice the model is manageable only when the effects are independent from each other, requiring many important assumptions. These include random mating, and hence Hardy-Weinberg equilibrium (i.e. no inbred individuals), linkage equilibrium (independent segregation of loci, which requires many generations to achieve for tightly linked genes) and no selection.

$$\begin{aligned} V_P &= V_G + V_E \\ &= V_A + V_D + V_I + V_E \end{aligned} \tag{3.4}$$

$$\begin{aligned} \sigma^2_P &= \sigma^2_G + \sigma^2_E \\ &= \sigma^2_A + \sigma^2_D + \sigma^2_I + \sigma^2_E \end{aligned} \tag{3.5}$$

Many more terms may be included, such as maternal genetic effects, and genotype × environment interaction. The model has unlimited opportunities for complexity. This is a strength, in that it is all-accommodating, and a weakness, in that datasets may allow to partition only a few components. In practice, assumptions must be made to reduce the complexity of the resemblance among relatives. Usually, the resemblance among relatives is assumed to depend only on additive genetic variance V_A and dominance variance V_D, so that the following sources of covariation are neglected:

- Epistatic variance (interaction effects among loci are small compared to additive and dominance effects)
- Environmental variance (effects of **shared environments** are assumed to be small enough)

3.1.3.2 Heritability:

The models and summary statistics defined by Fisher and Wright have remained at the heart of quantitative genetics, not least because they provide ways to make predictions of important quantities, such as

- Breeding value (A), the expected performance of of an individual's offspring
- Broad-sense heritability, the ratio of total genetic variance V_G to the overall phenotypic variance V_P:

$$
\begin{aligned}
H^2 &= V_G/V_P \\
&= (V_A + V_D + V_I)/V_P \\
H^2 &= \sigma_G^2/\sigma_P^2 \\
&= (\sigma_A^2 + \sigma_D^2 + \sigma_I^2)/\sigma_P^2
\end{aligned}
$$

- Narrow-sense heritability, the ratio of additive genetic variance V_A to the overall phenotypic variance V_P:

$$
\begin{aligned}
h^2 &= V_A/V_P \\
h^2 &= \sigma_A^2/\sigma_P^2
\end{aligned} \tag{3.6}
$$

- The response to artificial or natural selection, the increase (or decrease) of genetic values due to selection of individuals, over generations

In view of the assumed complexity of the underlying gene action, involving many loci with unknown effects and interactions, much quantitative genetic analysis has, unashamedly, been at a level of the 'black box'.

3.1.3.3 Genetic correlation:

In a general quantitative genetic model, in which, for each individual, two traits (P_1 and P_2) are are each defined as the sum of a genetic value (G_1 and G_2) and a environmental value (E_1 and E_2):

$$P_1 = G_1 + E_1 \tag{3.7}$$

$$P_2 = G_2 + E_2 \tag{3.8}$$

The phenotypic correlation ($\rho_{P_{12}}$) between the traits is defined as:

$$\rho_{P_{12}} = \frac{\sigma_{P_{12}}}{\sqrt{\sigma_{P_1}^2 \sigma_{P_2}^2}}$$

where $\sigma_{P_{12}}$ is the phenotypic covariance and $\sigma_{P_1}^2$ and $\sigma_{P_2}^2$ are the variances of the phenotypic values for the two traits in the population. The genetic correlation ($\rho_{G_{12}}$) of the traits is defined as:

$$\rho_{G_{12}} = \frac{\sigma_{G_{12}}}{\sqrt{\sigma_{G_1}^2 \sigma_{G_2}^2}}$$

where $\sigma_{G_{12}}$ is the genetic covariance and $\sigma_{G_1}^2$ and $\sigma_{G_2}^2$ are the variances of the genetic values for the two traits in the population.

3.1.4 Basic questions remain

On the premise that many genes and environmental factors interact to impact the trait, it will be difficult to determine the action of individual causal genes. Many basic questions remain: What do the genes do; how do they interact; on what traits does natural selection act; why is there so much genetic variation; and can we expect continued genetic improvement in selection programmes? Ultimately, we want to know at the molecular level not just which genes are involved, whether structural or regulatory, but what specific mutation (nucleotide substitution, deletious, copy number variant, etc.) is responsible for genetic effects, and how the causal genes are controlled.

Chapter 4

Estimation of Genetic Parameters

This section introduces the basic concepts of estimating genetic parameters such as:

- basic principles of estimating genetic parameters
- use of genetic relationships for estimating genetic parameters
- different methods, data sources and experimental designs for estimating genetic parameters
- importance of estimation of genetic parameters in breeding
- knowing when estimation of genetic parameters may be required

The estimation of genetic parameters is an important issue in animal and plant breeding. First of all, estimating additive genetic and possible non-additive genetic variances contributes to a better understanding of the genetic mechanism. Secondly, estimates of genetic and phenotypic variances and covariances are essential for the prediction of breeding values and for the prediction of the expected genetic response of selection programmes. Parameters that are of interest are heritability, genetic and phenotypic correlation and repeatability, and those are computed as functions of the variance components.

Genetic parameters are estimated using information on phenotypes and genetic relationships for individuals in the breeding population. Heritability is estimated by comparing individual phenotypic variation among related individuals in a population. Close (compared to distant) relatives share more DNA in common and if the trait is under genetic influence they will therefore share phenotypic similarities. In this section we will illustrate how different phenotypic sources and genetic relationships are used for estimating genetic parameters.

4.1 Genetic model

As introduced previously the phenotype for a quantitative trait is the sum of both genetic and environmental factors. In general the total genetic effect for an individual is the sum of both additive and non-additive effects. However, only the additive genetic effects are passed on to the offspring and therefore have a breeding value. Therefore we only consider the additive genetic model as the basis for estimation of genetic parameters. The model for the phenotype (y) is:

$$y = \mu + a + e$$

where μ is the population mean, a is the additive effect, and e is the environmental deviation (or residual) not explained by the genetic effects in the model. We assume that the additive genetic effect, a, and the residual term, e, are normally distributed which means that the observed phenotype is also normally distributed

$$a \sim N(0, \sigma_a^2), \quad e \sim N(0, \sigma_e^2), \quad y \sim N(\mu, \sigma_y^2)$$

where σ_a^2 is additive genetic variance, σ_e^2 is residual variance, and (σ_y^2) is the total phenotypic variance.

4.2 Genetic parameters

Heritability and genetic correlation are the key genetic parameters used in animal and plant breeding. They are defined in terms of the variance component (σ_a^2 and σ_e^2) defined in the previous section.

Heritability estimates the degree of variation in a phenotypic trait in a population that is due to genetic variation between individuals in that population. It measures how much of the variation of a trait can be attributed to variation of genetic factors, as opposed to variation of environmental factors. The narrow sense heritability is the ratio of additive genetic variance (σ_a^2) to the overall phenotypic variance ($\sigma_y^2 = \sigma_a^2 + \sigma_e^2$):

$$h^2 = \sigma_a^2/(\sigma_a^2 + \sigma_e^2) \tag{4.1}$$

A heritability of 0 implies that the no genetic effects influence the trait, while a heritability of 1 implies that all of the variation in the trait is explained by the genetic effects. In general the amount of information provided by the phenotype about the breeding value is determined by the narrow sense heritability.

Genetic correlation is the proportion of variance that two traits share due to genetic causes. Genetic correlations are not the same as heritability, as it is about the overlap between the two sets of influences and not their absolute magnitude; two traits could be both highly heritable but not be genetically

correlated or have small heritabilities and be completely correlated (as long as the heritabilities are non-zero). Genetic correlation (ρ_a) is the genetic covariance between two traits divided by the product of genetic standard deviation for each of the traits:

$$\rho_{a_{12}} = \frac{\sigma_{a_{12}}}{\sqrt{\sigma^2_{a_1}\sigma^2_{a_2}}} \tag{4.2}$$

where $\sigma_{a_{12}}$ is the genetic covariance and $\sigma^2_{a_1}$ and $\sigma^2_{a_2}$ are the variances of the additive genetic values for the two traits in the population. A genetic correlation of 0 implies that the genetic effects on one trait are independent of the other, while a correlation of 1 implies that all of the genetic influences on the two traits are identical. Thus in order to estimate the heritability and genetic correlation we need to estimate the variance component defined above.

4.3 Data required for estimating genetic parameters

Information on phenotypes and genetic relationships for individuals in the breeding population are used to formulate appropriate statistical models for the analysis of the data and accurate estimation of genetic parameters and breeding values of individuals.

Phenotypes for traits of economic importance need to be recorded accurately and completely. All individuals within a production unit (herd, flock, ranch, plot) should be recorded. Individuals should not be selectively recorded. Data includes the dates of events when traits are observed, factors that could influence an individual's performance, and an identification of contemporaries that are raised and observed in the same environment under the same management regime. Observations should be objectively measured, if at all possible.

Genetic relationships for the individuals in the breeding population is required. Genetic relationships can be inferred from a pedigree or alternative alternative from genetic markers. Individuals and their parents need to be uniquely identified in the data. Information about birth dates, breed composition, and genotypes for various markers or QTLs could also be stored. If individuals are not uniquely identified, then genetic change of the population may not be possible. In aquaculture species, for example, individual identification may not be feasible, but family identification (father and mother) may be known.

Furthermore knowledge and understanding of the production system is important for designing optimum selection and mating strategies. For dairy cattle key elements are the gestation length and the age at first breeding. The number of offspring per female per gestation will influence family structure. The use

of artificial insemination and/or embryo transfer could be important. Other management practices are also useful to know.

Prior information about the traits is useful. Read the literature. Most likely other researchers or breeders have already made analyses of the same species and traits. Their models could be useful starting points for further analyses. Their parameter estimates could predict the kinds of results that might be found. The idea is to avoid the pitfalls and problems that other researchers have already encountered. Be aware of new kinds of analyses of the same data, that may not involve statistical models.

4.4 Statistical models and variance components

For estimating genetic parameters we need to specify a model that describes the genetic and non-genetic factors that may affect the trait phenotypes. Often the non-genetic factors are referred to as systematic effect such as age, parity, litter size, days open, sex, herd, year, season, management, etc.

$$\text{mean} = \text{mean} + \text{systematic effect} + \text{genetic effect} + \textit{residual}$$

In this, we make a distinction between fixed effects, that determine the level (expected means) of observations, and random effects that determine variance. A model at least exists of one fixed (mean) and one random effect (residual error variance). If observations also are influenced by a genetic contribution of the individuals, then a genetic variance component exists as well. In that situation, we have two components contributing to the total variance of the observations: a genetic and a residual variance component.

A statistical model is usually specified as a mathematical relationship between one or more random variables and other non-random variables. In quantitative genetics it is often based on the additive genetic model specified above with inclusion of additional factors that may affect the trait of interest:

$$y = \mu + ...\text{systematic effect}.... + a + e$$

where μ is the population mean, a is the additive effect, and e is the environmental deviation (or residual) not explained by the systematic effects and the genetic effects in the model. We assume that the additive genetic effect, a, and the residual term, e, are normally distributed such that $a \sim N(0, \sigma_a^2)$ and $e \sim N(0, \sigma_e^2)$. The goal of the statistical analysis is to derive estimates for the variance components that is the additive genetic variance σ_a^2, and the residual variance σ_e^2 the residual variance.

The statistical model is a formal representation of our quantitative genetic theory, but it is important to realize that all models are simple approximations to how factors influence a trait. The goal of the statistical analysis is to find the best practical model that explains the most variation in the data. Statistical knowledge is required. The methods used for estimating genetic parameters is based on statistical concepts such as random variables, multivariate normal theory and linear (mixed) models. These concepts and their use will be explained in the following sections.

Chapter 5

Methods for estimation of genetic parameters

In general estimation of heritability and genetic correlation is based on methods that determine resemblance between genetically related individuals. Here we will present three methods for estimating heritability, parent-offspring regression, analysis of variance (ANOVA) for family data (e.g. halfsib/fullsib families) and restricted maximum likelihood (REML) analysis for general pedigree.

5.1 Estimating heritability using parent - offspring regression

The simplest method for estimation genetic parameters is based on regression analysis. Heritability may be estimated by comparing phenotypes for traits recorded in parent and offspring. Parent-offspring regression compares trait values in parents (y_p) to trait values in their offspring (y_o). Estimation of heritability is based on a linear regression model:

$$y_o = y_p b_{o|p} + e_o. \tag{5.1}$$

The slope of the regression line ($b_{o|p}$) approximates the heritability of the trait when offspring values are regressed against the average trait in the parents. If only one parent's value is used then heritability is twice the slope. In other words the expected value of the regression line is $b_{o|p} = 0.5h^2$ (or h2 when regression is on mid-parent mean).

To better understand this relationship consider a situation where we have collected phenotypes on a number of father-offspring families. From standard

regression theory the slope can be determined as:

$$b_{o|f} = \frac{Cov(y_f, y_o)}{Var(y_o)} \tag{5.2}$$

where $Cov(y_f, y_o)$ is the covariance between the phenotypes of the father and the offspring and $Var(y_o)$ is the variance of the offspring phenotypes.

The phenotypes for the father (y_f) and the offspring (y_o) can be expressed as:

$$\begin{aligned} y_f &= \mu + a_f + e_f \\ y_o &= \mu + 0.5a_m + 0.5a_f + a_{mendelian} + e_o \end{aligned}$$

where μ is the population mean, a_m and a_f are the additive genetic effect for the mother and the father, a_{mendelian} is the mendelian deviation in the offspring, and e_f and e_o are the residual effect for the father and the offspring.

The offspring get half of the genes from each parent and therefore the breeding value for the offspring is the average of the parents' breeding values plus the Mendelian deviation:

$$a_{\text{offspring}} = \frac{1}{2}a_{\text{father}} + \frac{1}{2}a_{\text{mother}} + a_{\text{mendelian}}$$

(a = additive genetic value = breeding value) The term $a_{mendelian}$ is necessary, because two fullsibs i and j both having parents $father$ and $mother$ receive different random samples of the set of parental alleles. Hence the breeding values a_i and a_j of halfsibs i and j are not going to be the same. Furthermore we assume that the breeding values are normally distributed:

$$\begin{aligned} a_{father} &\sim N(0, \sigma_a^2) \\ a_{mother} &\sim N(0, \sigma_a^2) \\ a_{mendelian} &\sim N(0, 0.5\sigma_a^2) \end{aligned} \tag{5.3}$$

Therefor an expression for the covariance between the phenotypes of the parent and the offspring can be derived as:

$$\begin{aligned} Cov(y_f, y_o) &= Cov(a_f + e_f, 0.5a_m + 0.5a_f + a_{mendelian} + e_o) \\ &= Cov(a_f, 0.5a_f) \\ &= 0.5Cov(a_f, a_f) \\ &= 0.5\sigma_a^2 \end{aligned}$$

and similar for the variance variance of the offspring phenotypes:

$$\begin{aligned} Var(y_o) &= Var(0.5a_m + 0.5a_f + 0.5a_{mendelian} + e_o) \\ &= Var(0.5a_m) + Var(0.5a_f) + Var(a_{mendelian}) + Var(e_o) \\ &= 0.25Var(a_m) + 0.25Var(a_f) + Var(a_{mendelian}) + Var(e_o) \\ &= 0.25\sigma_a^2 + 0.25\sigma_a^2 + 0.5\sigma_a^2 + \sigma_e^2 \\ &= \sigma_a^2 + \sigma_e^2 \end{aligned}$$

Therefore the expected value of the regression coefficient for a father-offspring analysis is:

$$\begin{aligned} b_{o|f} &= \frac{0.5\sigma_a^2}{\sigma_a^2 + \sigma_e^2} \\ &= 0.5\frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2} \\ &= 0.5h^2 \\ h^2 &= 2b_{o|f} \end{aligned}$$

Similar relationships can be derived for other types of parent-offspring regression analysis (mother-offspring and mid- parent-offspring). The heritability can be estimated from the regression coefficient based on:

$$\begin{aligned} h^2 &= 2b_{o|m} \quad \text{(mother-offspring regression)} \\ h^2 &= 2b_{o|f} \quad \text{(father-offspring regression)} \\ h^2 &= b_{o|mf} \quad \text{(mean parent-offspring regression)} \end{aligned}$$

Offspring-parent regression is not often used in practice. It requires data on 2 generations, and uses only this data. It is based on the genetic relationship between parent and offspring which equals 0.5 (i.e. offspring get half of the genes form its parent), but it is not possible to utilize genetic relationships among parents. However, the method is robust against selection of parents.

5.2 Estimating heritability using ANOVA for family data

Genetic parameters have been estimated for many years using analysis of variance (ANOVA). This method require that individuals can be assigned to groups with the same degree of genetic relationship for all members. Family structures considered most often are paternal half-sib groups or full-sib groups. In the case

of paternal half-sib group all offspring of one sire are treated as one group and offspring of different sires are allocated to different groups.

Estimation of heritability using ANOVA is based on a linear model. Consider a situation where we have phenotypic observation for multiple offspring for a number of families (halfsib or fullsib). A simple linear model for the phenotypic observation for the jth offspring in the ith famility include the population mean (μ) and a family effect (f_i):

$$y_{ij} = \mu + f_i + e_{ij} \tag{5.4}$$

Assume n observations, with n_f families, with n_o is the number of offspring per family.

The total sum of squares (SST) is the sum of each of the observations squared:

$$SST = \sum_{i=1}^{n_f} \sum_{j=1}^{n_o} y_{ij}^2 \tag{5.5}$$

where y_{ij} is an observation on the jth offspring in the ith family.

The mean sum of squares (SSM) is n times the mean squared:

$$SSM = n\bar{y}_{\cdot\cdot}^2 \tag{5.6}$$

The model sum of squares (SSA) due to a particular factor (e.g. the family effect) is therefore the sum over all observations of the estimated (family) effect in each observation squared (in balanced data this is the difference between the family group mean and the overall mean):

$$SSA = \sum_{i=1}^{n_f} (\bar{y}_{i\cdot} - \bar{y}_{-\cdot})^2 \tag{5.7}$$

Notice that the sum of squares for the main effect (SSA) is the sum of all the squared estimates of f_i, because in a balanced data set the estimate of f_i is equal to $(y_{i\cdot} - y_{\cdot\cdot})$. In a balanced data, it is rather simple to determine the expectations for each sum of squares, because the number of observations per class of f is constant (n_o).

The residual sum of squares (SSE) due to the residual (error) is the sum over all observations of the residual effect in each observation squared (this is the difference between the observation and its group mean):

$$SSE = \sum_{i=1}^{n_f} \sum_{j=1}^{n_o} y_{ij}^2 \tag{5.8}$$

The total sum of squares can be expressed as a sum of the components described above:

$$SST = SSM + SSA + SSE \tag{5.9}$$

In balanced data, it is rather simple to estimate variance components, by setting the "Mean Squares" equal to their expectations. Those expectations are linear functions of the variance components and is derived based on statistical theory (Expected value of a sum of squares). In this simple model we can calculate estimate of the residual variance components ($\hat{\sigma}_e^2$) as:

$$\hat{\sigma}_e^2 = SSE/(n - n_f) \tag{5.10}$$

and and estimate of the family variance ($\hat{\sigma}_f^2$) as:

$$\hat{\sigma}_f^2 = (SSA/(n_f - 1) - \hat{\sigma}_e^2)/n_o \tag{5.11}$$

Calculating the variance between groups, involves partitioning the sum of squared observations (SS) due to different sources of variation in the model of analysis, groups of relatives being one of them, and equating the corresponding mean squares. Mean squares are derived as the SS divided by the associated degrees of freedom, to their expectations.

Using ANOVA, the covariance among members of a family or group of relatives is usually determined as the variance component between groups. For example, in case of a half-sib family model, the variance between half-sib families σ_s^2 and variance within half-sib families σ_e^2. As shown earlier, the half-sib family variance $\sigma_s^2 = 0.25\sigma_a^2$ while the variance within half-sib families is $0.75\sigma_a^2 + \sigma_e^2$

Fullsib families

$$\begin{aligned}
y_m &= \mu + a_m + e_m \\
y_f &= \mu + a_f + e_f \\
y_{o1} &= \mu + 0.5a_m + 0.5a_f + 0.5a_{mendelian_1} + e_{o1} \\
y_{o2} &= \mu + 0.5a_m + 0.5a_f + 0.5a_{mendelian_2} + e_{o2}
\end{aligned}$$

$$\begin{aligned}
Cov(y_{o1}, y_{o2}) &= Cov(0.5a_m + 0.5a_f + a_{mendelian1} + e_{o1}, 0.5a_m + 0.5a_f + a_{mendelian2} + e_{o2}) \\
&= Cov(0.5a_f, 0.5a_f) + Cov(0.5a_m, 0.5a_m) \\
&= 0.25Cov(a_f, a_f) + 0.25Cov(a_m, a_m) \\
&= 0.25\sigma_a^2 + 0.25\sigma_a^2 \\
&= 0.5\sigma_a^2
\end{aligned}$$

Halfsib families

$$
\begin{aligned}
y_{m1} &= \mu + a_{m1} + e_{m1} \\
y_{m2} &= \mu + a_{m2} + e_{m2} \\
y_f &= \mu + a_f + e_f \\
y_{o1} &= \mu + 0.5a_{m1} + 0.5a_f + 0.5a_{mendelian1} + e_{o1} \\
y_{o2} &= \mu + 0.5a_{m2} + 0.5a_f + 0.5a_{mendelian2} + e_{o2}
\end{aligned}
$$

$$
\begin{aligned}
Cov(y_{o1}, y_{o2}) &= Cov(0.5a_{m1} + 0.5a_f + a_{mendelian1} + e_{o1}, 0.5a_{m2} + 0.5a_f + a_{mendelian2} + e_{o2}) \\
&= Cov(0.5a_f, 0.5a_f) + Cov(0.5a_{m1}, 0.5a_{m2}) \\
&= 0.25Cov(a_f, a_f) + 0 \\
&= 0.25\sigma_a^2 \\
&= 0.25\sigma_a^2
\end{aligned}
$$

Therefore the heritability can be estimated from the variance component based on:

$$
\begin{aligned}
\sigma_{hs}^2 &= 0.25\sigma_a^2 \quad \text{(halfsib families)} \\
\sigma_{fs}^2 &= 0.5\sigma_a^2 \quad \text{(halfsib families)} \\
h^2 &= \frac{4\sigma_{hs}^2}{4\sigma_{hs}^2 + \sigma_e^2} \quad \text{(halfsib families)} \\
h^2 &= \frac{2\sigma_{fs}^2}{2\sigma_{fs}^2 + \sigma_e^2} \quad \text{(fullsib families)}
\end{aligned}
\tag{5.12}
$$

Data arising from experimental designs used for estimating genetic parameters are usually not balanced (i.e. number of offspring varies across families). Methods analogous to the ANOVA have been developed for unbalanced data. However REML is nowadays the method of choice for variance component estimation.

The estimation of variance components (within and between family components). If the variation within families is large relative to differences between families, the trait must be lowly heritable. Variance components are attributed to specific effects. For example, the (paternal) half-sib variance is due to differences between sires. The variance component represents the sire variance, which is a quarter of the additive genetic variance.

Estimation of variance components is easier to generalise, and this method is generally used to estimate genetic parameters. The remainder of this chapter will therefore mostly deal with variance component estimation.

5.3 Estimating heritability using Restricted Maximum Likelihood

Genetic parameters are nowadays estimated using restricted maximum likelihood (REML) or Bayesian methods. This method allow for estimation of genetic parameters using phenotypic information for individuals from a general pedigree.REML is based on linear mixed model methodology and use a likelyhood approach for estimating genetic parameters.

5.3.1 Linear mixed model:

The linear mixed model contains the observation vector for the trait(s) of interest, the factors that explain how the observations came to be, and a residual effect that includes everything not explainable.

A matrix formulation of a general model equation is:

$$y = Xb + Za + e$$

where

y : is the vector of observed values of the trait,

b : is a vector of factors, collectively known as fixed effects,

a : is a vector of factors known as random effects,

e : is a vector of residual terms, also random,

X, Z : are known design matrices that relate the elements of b and a to their corresponding element in y.

The **observation vector** contains elements resulting from measurements, either subjective or objective, on the experimental units (usually animals) under study. The elements in the observation vector are random variables that have a multivariate distribution, and if the form of the distribution is known, then advantage should be taken of that knowledge. Usually y is assumed to have a multivariate normal distribution, but that is not always true. The elements of y should represent random samples of observations from some defined population. If the elements are not randomly sampled, then bias in the estimates of b and a can occur, which would lead to errors in ranking individuals.

A **continuous factor** is one that has an infinite-like range of possible values. For example, if the observation is the distance a rock can be thrown, then a continuous factor would be the weight of the rock. If the observation is the rate of growth, then a continuous factor would be the amount of feed eaten.

Discrete factors usually have classes or levels such as age at calving might have four levels (e.g. 20 to 24 months, 25 to 28 months, 29 to 32 months, and

33 months or greater). An analysis of milk yields of cows would depend on the age levels of the cows.

In the traditional "frequentist" approach, **fixed** and **random** factors need to be distinguished.

If the number of levels of a factor is small or limited to a fixed number, then that factor is usually fixed. If inferences about a factor are going to be limited to that set of levels, and to no others, then that factor is usually fixed. If a new sample of observations were made (a new experiment), and the same levels of a factor are in both samples, then the factor is usually fixed. If the levels of a factor were determined as a result of selection among possible available levels, then that factor should probably be a fixed factor. Regressions of a continuous factor are usually a fixed factor (but not always).

If the number of levels of a factor is large, then that factor can be a random factor. If the inferences about a factor are going to be made to an entire population of conceptual levels, then that factor can be a random factor. If the levels of a factor are a sample from an infinitely large population, then that factor is usually random. If a new sample of observations were made (a new experiment), and the levels were completely different between the two samples, then the factors if usually random.

Expectation and variance of variables in the model:

In the statistical model (specified above) the random effect (a and e) and the phenotypes (y) are considered to be random variables that are assumed to follow a multivariate normal distribution.
In general terms the expectations of these random variables are:

$$\begin{aligned} E(y) &= E(Xb) + E(Za) + E(e) \\ &= Xb + 0 + 0 \\ &= Xb \end{aligned}$$

and the variance-covariance matrices are:

$$\begin{aligned} Var(a) &= G \\ Var(e) &= R \\ Var(y) &= ZGZ' + R = V \end{aligned} \tag{5.13}$$

where G, R and V are square matrices of genetic, residual and phenotypic (co)variances among the individuals in the data set. If we assume

$$\begin{aligned} Var(a) &= G \\ &= A\sigma_a^2 \\ Var(e) &= R \\ &= I\sigma_e^2 \\ Var(y) &= ZGZ' + R = V \\ &= A\sigma_a^2 + I\sigma_e^2 \end{aligned}$$

Assumptions and limitations of the model:

The third part of a model includes items that are not apparent in parts 1 and 2. For example, information about the manner in which data were sampled or collected. Were the animals randomly selected or did they have to meet some minimum standards? Did the data arise from many environments, at random, or were the environments specially chosen? Examples will follow. A linear model is not complete unless all three parts of the model are present. Statistical procedures and strategies for data analysis are determined only after a complete model is in place.

5.3.2 Likelihood approach for estimating variance components

Restricted Maximum Likelihood is a method that is used to estimate the parameters in the model (i.e. variance components σ_a^2 and σ_e^2) specified in the linear mixed model above. The general principle used in likelihood methods is to find the set of parameters which maximizes the likelihood of the data.

It is useful to recall that the likelihood ($L(\theta|y)$) is any function of the parameter (θ) that is proportional to $p(y|\theta)$. Maximizing $L(\theta|y)$ leads to obtaining the most likely value of θ ($\widehat{\theta}$) given the data y. Usually the likelihood is expressed in terms of its logarithm ($l(\theta|\mathbf{y})$) as it makes the algebra easier.

The likelihood of the data for a given linear mixed model can be written as a function;

$$l(\mathbf{V}|\mathbf{y}, \mathbf{X}, \beta) \propto -\frac{1}{2}\ln|\mathbf{V}| - \frac{1}{2}\ln|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\beta)'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta) \quad (5.14)$$

From calculus we know that we can find the maximum of a function by taking the first derivative and set that equal to zero. Solving that would result in the desired parameters (assuming that we did not find the minimum, this can

be checked using second derivatives). The first and second derivatives of the likelihood function are complicated formulas.

There are no simple one-step solutions for estimating the variance components based on REML (**?**). Instead, we infer the partial derivatives of the likelihoods with respect to the variance components. The solutions to these involve the inverse of the variance-covariance matrix, which themselves includes the variance components, so the variance components estimates are non-linear functions of the variance components. It is therefore necessary to apply iterative methods to obtain the estimates.

From the estimate of the variance components the heritability can easily computed by

$$\hat{h}^2 = \hat{\sigma}_a^2/(\hat{\sigma}_a^2 + \hat{\sigma}_e^2) \tag{5.15}$$

where the "$\hat{h}^2$" refers to the heritability is an estimate.

5.3.3 Advantages of using REML for estimating genetic parameters

The REML method was developed by **?** as an improvement of the standard Maximum Likelihood (ML). The ML method was originally proposed by **?** but was introduced to variance components estimation by **?**. ML assumes that fixed effects are known without error which is in most cases false and, as consequence, it produces biased estimates of variance components (usually, the residual variance is biased downward). To solve this problem, REML estimators maximize only the part of the likelihood not depending on the fixed effects, by assuming that the fixed effects have been, so to speak, fixed. This entails that when comparing multiple models by their REML likelihoods, they must contain the same fixed effects, and that REML, by itself, does not estimate the fixed effects.

REML does not produce unbiased estimates owing to the inability to return negative values of variance components of many methods to obtain REML estimators, but it is still the method of choice due to the fact that this source of bias is also present in ML estimates (**?**).

REML requires that y have a multivariate normal distribution although various authors have indicated that ML or REML estimators may be an appropriate choice even if normality does not hold (Meyer, 1990).

REML can account for selection when the complete mixed model is used with all genetic relationships and all data used for selection included (Sorensen and Kennedy, 1984; Van der Werf and De Boer, 1990).

There is obviously an advantage in using (RE)ML methods that are more flexible in handling animal and plant breeding data on several (overlapping) generations (and possibly several random effects). However, the use of such methods has

a danger in the sense that we need not to think explicitly anymore about data structure. To estimate, as an example, additive genetic variance, we need to have a data set that contains a certain family structure that allows us to separate differences between families from differences within families. Or in other words, we need to separate genetic and residual variance. ANOVA methods require more explicit knowledge about such structure, since the data has to be ordered according to family structures (e.g. by half sib groups).

Developments in variance component estimation specific to animal and plant breeding have been closely linked with advances in the genetic evaluation using Best Linear Unbiased Prediction (BLUP). Early REML applications were generally limited to models largely equivalent to those in corresponding ANOVA type analysis, considering one random effect only and estimating genetic variances from paternal half sib covariances (so-called sire model). Today, heritability can be estimated from general pedigrees and from genomic relatedness estimated from genetic markers. Linear mixed models is also used genetic evaluation schemes, allowing information on all known relationships between individuals to be incorporated in the analysis. Linear mixed models can include maternal, permanent environmental, cytoplasmic or dominance effects or effects at QTL thereby more accurately describe the observed data. These effects are fitted as additional random effects.

5.4 When to estimate variance components?

In general, the estimation of variances and covariances has to be based on a sufficient amount of data. Depending on the data structure and the circumstances during measuring, estimations can be based on some hundreds (selection experiments) or more than 10,000 observations (field recorded data). It is obvious that we are not interested in estimating variance components from every data set. The information in literature is in many cases even better than estimations based on a small data set. In general, we have to estimate variance if: - we are interested in a new trait, from which no parameters are available; - variances and covariances might have changed over time - considerable changes have occurred in a population e.g. due to recent importations. Mostly it is assumed that variances and covariances, and especially the ratio of both of them (like heritability, correlation), are based on particular biological rules, which do not rapidly change over time. However, it is well known that the genetic variance changes as consequence of selection. Changes are especially expected in situations with short generation intervals, high selection intensities or high degrees of inbreeding or in a situation in which a trait is determined by only a few genes. Secondly, the circumstances under which measurements are taken can change. If conditions are getting more uniform over time, the environmental variance decreases, and consequently the heritability increases. Thirdly, the biological interpretation of a trait can change as consequence of a changed environment; feed intake under limited feeding is not the same as feed intake under ad-lib

feeding. In conclusion, there are sufficient reasons for regular estimation of (co-)variance components.

Chapter 6

Applications

Some *significant* applications are demonstrated in this chapter.

6.1 Example one

6.2 Example two

Chapter 7

Final Words

We have finished a nice book.

Bibliography

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2020). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.20.