SIMULATION FOR THE INVESTIGATION OF A SOCIAL MODEL FOR INDIRECT GENETIC EFFECTS IN AQUACULTURE

by

R. Benjamin Dexter

Submitted in partial fulfillment of the
requirements for the degree of
Master of Science

at

Dalhousie University
Halifax, Nova Scotia
December 2015

Table of Contents

List of Tables

List of Figures

Abstract

Aquaculture is growing quickly as a method for farming aquatic based organisms for use in human consumption, and it is likely that it will produce over 60% of total fish used in human consumption by 2030. This growth will require productivity increases. Unlike land animals fish raised in a farming environment are in very tight quarters resulting in competitive interactions. An important question is how to model this indirect genetic effect and how well it can be estimated compared to well understood direct genetic effects. An experiment with fish was performed at Dalhousie University to investigate a model including direct and indirect genetic effects. This work investigates the model proposed by Peter Bijma, explaining its background statistically and biologically and then performing a simulation study to determine how well this model behaves under various circumstances, namely different possible values of parameter sets for the model, and to see how well it estimates the variances involved in a situation deemed optimal by the model proposer. Interesting results are seen in that the modelling of the covariance is more complicated than expected, and that the direct genetic variance is more easily accurately estimated than the indirect genetic variance, as would be expected in a biological context.

Acknowledgements

I would like to thank my father for supporting me while I pursued my schooling, and for pushing me to continue with it as I saw fit. I would also like to thank Dr. Bruce Smith for providing an environment I worked well in, and helping me see that what seemed like an impossible amount of work was accomplishable. I would thank Dr. Christophe Herbinger and Dr. Joanna Flemming for the input and comments they provided to me in the final stages of writing. Finally I would express my gratitude to Phillipe Fullsack for his instrumental help in the early phases of this project, and getting it off the ground.

Chapter 1

Introduction

This work is primarily focusing on the breeding of fish in an aquaculture environment, and statistical analysis of animal breeding in this context. In particular it was designed to accompany a particular experiment occurring at Dalhouse University in order to augment the understanding of the results of this experiment. The reason for this is that the experiment plans to use a particular model in order to analyse the data gathered, which has been recently proposed in a theoretical sense, and not used practically in this environment to a large degree. The remainder of this chapter will briefly outline aquaculture, its importance and prevalence in Canada and globally. It will also present the basics of the experiment preformed in order to familiarise the reader with the context in which the simulation and analysis will occur.

1.1 Aquaculture Introduction

Aquaculture is the cultivation or farming of aquatic organisms, either in fresh or salt water, such as fish, molluscs, plants etc. under controlled conditions, similar to what agriculture is for land based organisms. Fish in particular are raised commercially in tanks or ponds, generally for use as food. This is a fast growing industry in food production and accounted for about 48.9% of the global fish food used in human consumption in 2012 (FAO 2014 [9]). It is likely that this industry will continue to expand, and it is expected that more than half the total seafood production worldwide will be produced by aquaculture in 2030, with aquaculture producing 60% of all fish intended for direct human consumption (World Bank 2014 [32]).

This increase in aquaculture production will require increases in productivity throughout the industry. In order to achieve this, aquaculture will need to take cues from the agriculture industry, and rely on genetically superior stock subjected to breeding programs. This will allow more fish to be produced, more will survive to harvesting age etc. In Canada most of the aquaculture is based on wild or poorly

domesticated stocks, with a few exceptions, which is unusual compared to the rest of common farming activity.

Animals involved in aquaculture are often raised in dense groups, and this can result in strong competitive interactions between individuals in the group. In particular social interactions can have strong effects on growth or survival. This can complicate quantitative genetic parameter estimation due to the extra interaction. Negative interactions are known to occur in tilapia and carp, which show lower than normal growth in the presence of larger animals, and similar interactions are suspected in Arctic Char and other species (Jobling 1983 [21]). Thus, selection for higher growth rate may lead to also selecting for more aggressive behaviour, which could slow down, or even negate any gains made from size selection. This can be seen in terrestrial animal breeding programs where animals compete for resources and it has been surmised that the presence of these sorts of social interactions can explain why some behavioural traits fail to respond to selection (Bijma et al 2007 [3]). Competition and interactions between group members may be a very strong source of size variation in fish or crustaceans raised in groups (Karplus 2005 [22]) and not accounting for this may be a reason for poor performance in aquaculture breeding programs to date (Jobling 1983 [21]).

I will use Linear Mixed Models in order to model social effects such as those proposed by Muir (2005) [33] and Bijma et al (2007) [3]. Social effects have the potential to strongly influence the response to selection in fish. In classical terms, the genetic value of an individual or breeding value, does not take into account relationships with other individuals, or social interactions. A new model has been proposed by Muir and Bijma that recognises that the genotype of an individual (the particular alleles that an individual carries), is not the only factor contributing toward the phenotype of an individual (outward expression of a particular trait), but also the fact that the individual is located in a group of other individuals. The model includes additional effects accounting for this "social group", leading to a new partition to the phenotypic variance and to a new definition of the breeding values of individuals. The variance of this extended breeding value could be much larger than the variance of standard breeding values in some circumstances. It has been shown in quail (Muir 2005 [33]) and pigs (Bergsma et al 2008[1]) that social interactions contribute the

majority of heritable variance in growth. This has not been explored deeply in aquaculture breeding, and will be explored. In particular this model proposed by Bijma will be investigated using simulations in order to determine how well it preforms under varying circumstances.

1.2 Experiment Details

An experiment was planned and carried out at Dalhousie University in order to investigate the social effects on growth in Arctic Char. While the present work focuses on the particular model proposed by Muir and Bijma in order to investigate how it behaves, this was done in conjunction with this experiment such that the results of the experiment could have a baseline to work with, as this type of model has not been applied to a real aquaculture setting previously. For this reason, the simulations carried out for this work were performed with this particular social experiment in mind, along with conclusions from Bijma about optimal designs for estimation.

Arctic Char were brought to Dalhousie's Aquatron facilities in Halifax, N.S, as fertilized eggs. These facilities are located on Dalhousie's main campus, and contain several very large tanks, as well as many small tanks and research equipment. These fish were from 24 families, some being half-sib (one shared parent), and others being full-sib (both parents shared). There were 10 females and 24 males that were used to create these families, with each male only ever being used once in family creation. Each of these families contained about 300-400 eggs when they arrived at Dalhousie. These eggs were then placed in tanks labelled by family and allowed to develop normally. See Figure 1.1.

Determining how to distribute the fish in an experimental design was challenging, as no experimental trials with fish exist that have followed the social and direct effects model proposed. There have been some involving land animals, namely with chickens (Craig & Muir 1996 [5]) or pigs (Bergsma et al., 2008 [1]). Performing an experiment of this nature with land animals allows for a very large number of isolated rearing units in which to raise the animals. In the previously mentioned trials, hundreds or thousands of different groups can be used, each with only a small number of individuals. In the case of Craig and Muir, each group contained only 2 families, and used a very large number of groups to allow different families as much exposure to

Figure 1.1: The tanks used to hatch the eggs in the Aquatron Facility

others as possible. Whereas with fish the isolated groups must be tanks, and with only 24 families of Arctic Char 276 tanks would be required, each with only a few fish in order to make a completely balanced design. This is not easily accomplished in an experimental setting, nor is it representative of realities in the aquaculture industry.

Thus the design chosen to be used was to place 10 individuals from 5 non-related families together in a tank, giving a group size of 50 individuals. These fish were all marked by family in order to be able to determine which fish belonged to which family. This was repeated in 24 tanks total, following a pattern of cyclical allocation in order to ensure both that every family was used equally, and that each was exposed to as many as possible. As an example, tank 1 would contain the unique families 1,2,3,4 and 5, while tank 2 contained families 2,3,4,5, and 6 etc. This culminates in tank 24 containing families 24,1,2,3, and 4. This design can be seen in Table 1.1. This gives us 24 tanks, with 5 Families/Tank, 10 Fish/Family, giving 50 Fish/Tank with 1200 fish total. There were sufficient fish from all of the families in order to create a replicate of this design in another set of 24 different tanks, and remaining fish were used to create a lower density third design not in the scope of this work, or held in tanks to keep them separate from the experiments. After a fixed periods of time the sizes of the fish in the tanks were measured, and these measurements represent the outcome of the experiment.

The remainder of the this thesis is as follows. Chapter 2 reviews estimation for general linear mixed models, inducing a discussion of of the animal model used in the estimation of breeding values. Chapter 3 introduces the concept of indirect genetic, or "social" effects, and describes Bijma's model for including direct and indirect effects in the model for estimating breeding values. The remainder of chapter 3 describes some results of an extensive simulation study used to assess the ability to reliably estimate direct and indirect genetic effects, with focus on the variances and covariances of direct and indirect components. Chapter 4 provides some general conclusions regarding the methods used and the ability to estimate indirect genetic effects together with some suggestions for future work.

Tank	**Families**				
1	1	2	3	4	5
2	2	3	4	5	6
3	3	4	5	6	7
4	4	5	6	7	8
5	5	6	7	8	9
6	6	7	8	9	10
7	7	8	9	10	11
8	8	9	10	11	12
9	9	10	11	12	13
10	10	11	12	13	14
11	11	12	13	14	15
12	12	13	14	15	16
13	13	14	15	16	17
14	14	15	16	17	18
15	15	16	17	18	19
16	16	17	18	19	20
17	17	18	19	20	21
18	18	19	20	21	22
19	19	20	21	22	23
20	20	21	22	23	24
21	21	22	23	24	1
22	22	23	24	1	2
23	23	24	1	2	3
24	24	1	2	3	4

Table 1.1: Experimental Design - Family Distribution

Chapter 2

Best Linear Unbiased Prediction

Statistical models that are used in this work, both in the background material, as well as the main matter all make use of both fixed and random effects. The most common tool for dealing with these models is BLUP, or Best Linear Unbiased Prediction. Knowledge of how this tool is used, and how it works is important in understanding the basic models that are used in the genetic context, as well as the more complicated versions, and the particular model that we will be investigating. These basic models are also of relevance, as they provide the framework with which to move forward in a genetic context, and provide a large amount of information about sources of variance, and how genetic relations between animals influence estimation. This chapter will focus on providing the statistical and genetic background material of relevance to the environment the simulation study will be taking place in.

2.1 Introduction

Best Linear Unbiased Prediction, or BLUP is a general method of dealing with estimation and prediction problems involving both fixed and random effects. It is in particular used in genetic estimation problems, typically in the context of estimating breeding values (a random effect), or other means of determining the influence of genetics when dealing with breeding programs or experiments.

The following model is known as the General Mixed Model, as it is a model that takes into account both fixed and random effects. The matrix form of this model is,

$$y = X\beta + Zu + e$$

where y is a column vector containing n observable random variables, or responses. In the context of genetic analysis, this would be a vector of phenotypic values for a trait in each of n individuals.

We then assume that these values are described by a linear model with fixed effects and random effects. These are represented by β a $p \times 1$ vector of unknown parameters having fixed values, and u a $q \times 1$ vector of unobservable random variables.

X and Z are known matrices of sizes $n \times p$ and $n \times q$ respectively. X is often known as a Design Matrix, as it is the matrix cataloguing which observation was influenced by the different fixed effects in the model, generally determined by experimental design. The elements in these matrices are often values of 0 or 1. In the genetic context, this is indexing which individuals are being influenced by particular random or fixed effects.

Finally e is the $n \times 1$ column vector of residuals, which are assumed to be distributed independently from the random effects such that $E[u] = 0$, $E[e] = 0$ This gives $E[y] = X\beta$ and

$$\text{Var} \begin{bmatrix} u \\ e \end{bmatrix} = \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix} \sigma^2$$

with the $p \times p$ matrix G being the covariance matrix for the vector of random effects u and the $n \times n$ matrix R being the covariance matrix for the vector of residual errors e. These are both positive definite matrices and σ^2 a positive constant. Under the assumption that u and e are uncorrelated, the covariance matrix for the vector of responses y is

$$V = ZGZ^T + R$$

ZGZ^T denotes the contribution to variance from the random effects, and R is denoting the contribution from residual error effects. In general it is assumed that residual errors are uncorrelated and of constant variance. Under these assumptions R becomes a diagonal matrix of the form $R = \sigma_E^2 I$.

2.2 Estimation

Often the goal with performing estimation when in the genetic framework is to estimate variance components. However in general inferences about the fixed effects

or random effects are often desired. In particular in a genetic analysis these may be such things as breeding value of an individual (random), or effect of a particular environment (fixed). In general estimators of random effects are known as predictors, in order to recognise the difference between fixed and random effects estimators. In this sense BLUP is a best linear unbiased predictor in that $E[BLUP(u)] = u$ satisfying the unbiased piece, linear in that it deals with linear functions of y, and best in that it minimises sampling variance. Whereas best linear unbiased estimators (BLUE) follow the same pattern, however dealing with the fixed effects, such that $E[BLUE(\beta)] = \beta$.

The BLUE for β is the standard generalised least squares estimator as below:

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y$$

and the BLUP of u is as follows (Henderson 1963) [16]

$$\hat{u} = GZ^T V^{-1}(y - X\hat{\beta})$$

It is noteworthy that both the estimate of $\hat{\beta}$ and $\hat{u}$ above require the inversion of the $V = ZGZ^T + R$ matrix. This is not trivial, as V can contain a very large number of entries if the y vector contains a large number of observations. This can make the computation of V^{-1} and therefore $\hat{\beta}$ and $\hat{u}$ difficult. As an answer to this problem, Henderson(1950) [15] developed a different way to compute $\hat{\beta}$ and $\hat{u}$ jointly in his mixed model equations, derived as follows:

Assume that u and e are normally distributed such that $u \sim (0, G)$, $e \sim (0, R)$ and $cov(u, e) = 0$. We then wish to maximise $f(y, u)$, the joint density of y and u with respect to β and u.

$$\begin{aligned} f(y,u) = (2\pi\sigma^2)^{-\frac{1}{2}n-\frac{1}{2}q} \left(det \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix} \right)^{-\frac{1}{2}} \\ \cdot exp \left\{ -\frac{1}{2\sigma^2} \begin{pmatrix} u \\ y - X\beta - Zu \end{pmatrix}^T \cdot \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}^{-1} \begin{pmatrix} u \\ y - X\beta - Zu \end{pmatrix} \right\} \end{aligned} \tag{2.2.1}$$

Maximizing this with respect to β and u means we need to minimise the quadratic form

$$\begin{pmatrix} u \\ y - X\beta - Zu \end{pmatrix}^T \cdot \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}^{-1} \begin{pmatrix} u \\ y - X\beta - Zu \end{pmatrix}$$
$$= \mathrm{u}^T G^{-1} u + (y - X\beta - Zu)^T R^{-1} (y - X\beta - Zu).$$

Differentiating this with respect to β and u using the rules for vector derivation of scalar functions gives the following

$$\frac{\delta f}{\delta \beta} = -X^T R^{-1}(y - X\beta - Zu)$$
$$\frac{\delta f}{\delta u} = 2G^{-1}u - 2Z^T R^{-1}(y - X\beta - Zu)$$

Then setting these derivatives to 0 and solving allows obtaining the following equations

$$X^T R^{-1} y = X^T R^{-1} X\hat{\beta} + X^T R^{-1} Z\hat{u} \tag{2.2.2}$$

$$Z^T R^{-1} y = Z^T R^{-1} X\hat{\beta} + (Z^T R^{-1} Z + G^{-1})\hat{u} \tag{2.2.3}$$

These equations are known as the "Mixed Model Equations" or MME's. Note that they do not contain V^{-1} as they instead use the R and G matrices. These tend to be diagonal, and thus R^{-1} and G^{-1} are much easier to obtain than V^{-1}. The same equations are occasionally written in a matrix form as below

$$\begin{pmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{u} \end{pmatrix} = \begin{pmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{pmatrix} \tag{2.2.4}$$

It is worth noting that even with these MME's, the application of either these equations, or the estimators given above, the variance components are a required element. This means that for practical application of BLUP the variance components need to be estimated as well by methods such as ANOVA or REML, which will be discussed later. In the particular case of genetics and the animal model that we investigate, the variance components are still required, however they are simplified as a lot of the variance structure is present in what is known as the "Relationship Matrix" to be discussed later.

2.3 Standard Errors

Looking at the matrix form of the mixed model equations above, if we allow the inverse of the leftmost matrix to be the following:

$$\begin{pmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{pmatrix}^{-1} = \begin{pmatrix} C_{11} & C_{12} \\ C_{12}^T & C_{22} \end{pmatrix} \tag{2.3.1}$$

Using this notation, Henderson(1975) [17] showed that the sampling covariance for the BLUE of $\hat{\beta}$ is

$$\sigma(\hat{\beta}) = C_{11}$$

the sampling covariance for the prediction errors $(\hat{u} - u)$ is given as

$$\sigma(\hat{u} - u) = C_{22}$$

and that the sampling covariance is

$$\sigma(\hat{\beta}, \hat{u} - u) = C_{12}$$

whereas the standard errors of the fixed and random effects are the square roots of the diagonal elements of C_{11} and C_{22} respectively. For large designs, the inverse of the matrix on the left hand side of 2.3.1 may be non-trivial, especially in the case of animal breeding where this matrix can have a large number of elements due to the number of animals involved. Due to this, the Mixed Model Equations are generally solved iteratively in order to determine the estimates of $\hat{\beta}$ and $\hat{u}$. This avoids inverting the coefficient matrix, however it also does not provide the diagonal elements of this inverted matrix, so the standard errors are not readily obtained.

Meyer(1989a) [27] showed a method of approximating the diagonal elements of this matrix in the particular case of animal breeding in order to estimate the standard errors in a way that doesn't involve the inversion of the coefficient matrix. The method used is to first isolate individual i's portion of the matrix, with the i^{th} diagonal element up to the $(i+3)^{th}$ element. This includes the individual, its parents, the number of records in the subclass of fixed effects that the i^{th} individual belongs to, and a value

representing a heritability in the population. Then partitioned matrix results are used to invert this submatrix's elements. If this submatrix was the entire matrix of interest, or that the remainder of the rows and columns were not connected to individual i, than this would yield true values. Meyer does this procedure for each of the individuals , and then preforms an adjustment on each of the estimated diagonal values, which adjusts parents by their progeny, progeny by how much information is known about their parents, namely if both sire and damn are known or not, and by the number of animals associated with each fixed effect group. Meyer used simulations to show that this estimation does well, and that the adjustment done makes a large difference in the accuracy of the estimate under certain conditions, but in general the estimate is consistently larger than the true values, and so Meyer suggests a scaling factor should be included as well.

2.4 Bayesian Derivation

In order to derive BLUP in a Bayesian context, take β as a parameter with a uniform improper prior distribution and u as a parameter with a prior distribution with mean 0 and variance $G\sigma^2$ independent of β. Given these two parameters the density of y is then

$$(2\pi\sigma^2)^{-\frac{1}{2}n} det(R)^{-\frac{1}{2}} exp - \tfrac{1}{2\sigma^2}(y - X\beta - Zu)^T R^{-1} \cdot (y - X\beta - Zu)$$

The prior density is

$$(2\pi\sigma^2)^{-\frac{1}{2}q} det(G)^{-\frac{1}{2}} exp - \tfrac{1}{2\sigma^2}(u^T G^{-1} u)$$

This shows that the posterior density for β and u is proportional to the joint density shown earlier (2.2.1), and therefore the posterior mode is given by the BLUP estimates.

2.5 Goldberger

Goldberger(1962) [11] is often credited with the first use of the term "Best Linear Unbiased Predictor". Goldberger gives his own derivation of BLUP which is rather straightforward, however he is following a slightly different starting model. Golberger considers the model:

$$y = X\beta + \epsilon \tag{2.5.1}$$

$$E[\epsilon] = 0 \tag{2.5.2}$$

$$E[\epsilon\epsilon^T] = \Omega \tag{2.5.3}$$

He then considers the problem of predicting a single drawing of response given a vector of the regressor variables. This drawing is written as

$$y_* = x^T\beta + \epsilon_* \tag{2.5.4}$$

Note that this model Goldberger is examining does not include u, our vector of random effects. However, Goldberger investigates this model under the view that Ω is not proportional to the identity matrix, and therefore it is not reasonable to assume that the prediction error (ϵ_*) is independent of the sample error. This sufficiently complicates the variance structure to allow us to equate the two situations, where in Goldbergers case our random effects are included in his error structure. This view that the errors are not independent leads Goldberger to starting with the following assumptions

$$E[\epsilon_*] = 0 \tag{2.5.5}$$

$$E[\epsilon_*^2] = \sigma_*^2 \tag{2.5.6}$$

$$E[\epsilon_*\epsilon] = w \tag{2.5.7}$$

Starting from these assumptions we then wish to to find the best linear unbiased predictor of y_*. In other words, with c being a vector of constants we wish to find the linear predictor

$$p = c^T y \tag{2.5.8}$$

Such that

$$\sigma_p^2 = E(p - y_*)(p - y_*)^T$$

$$\text{is minimised subject to } E(p - y_*) = 0 (2.5.9)$$

From 2.5.1,2.5.4,2.5.8 we obtain

$$p = c^T X\beta + c^T\epsilon \tag{2.5.10}$$

$$p - y_* = (c^T X - x_*^T)\beta + c^T\epsilon - \epsilon_* \tag{2.5.11}$$

For this to be unbiased, $E[p - y_*] = 0$ it requires that

$$c^T X - x_*^T = 0 \tag{2.5.12}$$

Then for an unbiased prediction

$$p - y_* = c^T\epsilon - \epsilon_* \tag{2.5.13}$$

and the prediction variance is

$$\sigma_2^2 = E(p - y_*)(p - y_*)^T = E[c^T\epsilon\epsilon^T c + \epsilon_*^2 - 2c^T\epsilon_*\epsilon]\sigma_p^2 = c^T\Omega c + \sigma_*^2 - 2c^T w \tag{2.5.14}$$

using 2.5.3, 2.5.6, and 2.5.7.

Then to minimise 2.5.14 subject to 2.5.12 we can minimise

$$d = c^T\Omega c - 2c^T w - 2\lambda^T(X^T c - x_*) \tag{2.5.15}$$

with λ being a vector of Lagrangian multipliers. If then we differentiate d with respect to c and λ we obtain:

$$\frac{d}{\delta c} = 2\Omega c - 2w - 2X\lambda \tag{2.5.16}$$

$$\frac{\delta d}{\delta \lambda} = 2X^T c - 2x_* \tag{2.5.17}$$

We then set the above two equations to 0 in order to solve for the minimising values of c ($\hat{c}$) and λ ($\hat{\lambda}$). This gives

$$\begin{bmatrix} \Omega & X \\ X^T & 0 \end{bmatrix} \begin{bmatrix} \hat{c} \\ -\hat{\lambda} \end{bmatrix} = \begin{bmatrix} w \\ x_* \end{bmatrix} \tag{2.5.18}$$

Applying the rule for inversion of a partitioned matrix gives the solution to 2.5.18

$$\begin{bmatrix} \hat{c} \\ \hat{\lambda} \end{bmatrix} = \begin{bmatrix} \Omega^{-1}[I - X(X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1}] & \Omega^{-1}X(X^T\Omega^{-1}X)^{-1} \\ (X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1} & -(X^T\Omega^{-1}X)^{-1} \end{bmatrix} \begin{bmatrix} w \\ x_* \end{bmatrix} \tag{2.5.19}$$

this gives the solution to $\hat{c}$ as follows

$$\hat{c} = \Omega^{-1}X(X^T\Omega^{-1}X)^{-1}x_* + \Omega^{-1}[I - X(X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1}]w \tag{2.5.20}$$

Thus we have the best linear unbiased predictor

$$\hat{p} = \hat{c}y = x_*^T(X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1}y + w^T\Omega^{-1}y - w^T\Omega^{-1}X(X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1}y \tag{2.5.21}$$

Translating Goldberger's work to the notation and model we've been using so far makes the following substitutions.

$$\epsilon = Zu + e$$

and

$$\Omega = (ZGZ^T + R)\sigma^2$$

To estimate $x_*^T\beta + z_*^T u$ take

$$\epsilon_* = z_*^T u$$

So we then have

$$w^T = E[z_*^T u(Zu + e)^T] = z_*^T GZ^T\sigma^2$$

This means that based on Goldberger's derivation above, the best linear unbiased predictor of $x_*^T\beta + z_*^T u$ is

$$x_*^T[X^T(ZGZ^T + R)^{-1}]^{-1}X^T(ZGZ^T + R)^{-1}y + z_*^T GZ^T(ZGZ^T + R)^{-1}y$$
$$\text{- } z_*^T GZ^T(ZGZ^T + R)^{-1}X[X^T(ZGZ^T + R)^{-1}X]^{-1}X^T(ZGZ^T + R)^{-1}y$$

Note since $V = ZGZ^T R$ that
$x_*^T[X^T(ZGZ^T + R)^{-1}]^{-1}X^T(ZGZ^T + R)^{-1}y = \hat{\beta}$ the above reduces to

$$x_*^T\hat{\beta} + z_*^T GZ^T(ZGZ^T + R)^{-1}(y - Z\hat{\beta}) \tag{2.5.22}$$

A common matrix identity used in this subject matter is

$$(Z^T R^{-1}Z + G^{-1})^{-1}Z^T R^{-1} = GZ^T(R + ZGZ^T)^{-1}$$

Applying this to 2.5.22 gives

$$x_*^T\hat{\beta} + z_*^T(Z^T R^{-1}Z + G^{-1})^{-1}Z^T R^{-1}(y - X\hat{\beta}) \tag{2.5.23}$$

Recall the second of the Mixed Model Equations (2.2.3)

$$Z^T R^{-1}y = Z^T R^{-1}X\hat{\beta} + (Z^T R^{-1}Z + G^{-1})\hat{u}$$

If we solve this for $\hat{u}$ we obtain

$$\hat{u} = (Z^T R^{-1} Z + G^{-1})^{-1} Z^T R^{-1} (y - X\hat{\beta}) \tag{2.5.24}$$

It is then easy to see that 2.5.23 and 2.5.24 combined produce

$$x_*^T \hat{\beta} + z_*^T \hat{u}$$

Showing that the predictor Goldberger derived is equivalent to that of the Mixed Model Equations. A side note is that if we take 2.5.24 and apply the matrix identity above to it, as well as substitute $(R + ZGZ^T) = V$ we get

$$\hat{u} = GZ^T V^{-1} (y - X\hat{\beta})$$

which is the BLUP of u as stated at the beginning of the chapter.

2.6 Animal Model

The topic of genetics and animal breeding often deals with the concept of a "Breeding Value".This is the sum of the average effects of alleles of a breeding animal and is measured based on the performance of its offspring. In other words, how valuable an individual's genes are for producing desired traits in offspring, based on measuring the offspring.

The animal model is used to estimate breeding values of measured individuals. Other common models are the Gametic model, a variation of the animal model where the breeding values are measured in terms of parental contribution, and the Reduced Animal Model, which combines the two for use in specific cases where parental breeding values are the topic of interest.

The animal model is a particular case of the General Mixed Model described earlier. The simplest version is the case with only one fixed factor, the population mean. In this case individual i's observation is expressed as the following:

$$y_i = \mu + a_i + e_i$$

Where μ is the population mean, and a_i is the additive genetic value, or breeding value, of the i^{th} individual. This is just a particular case of the General Mixed Model with

$$X = \begin{pmatrix} 1 \\ 1 \\ \vdots \end{pmatrix}, \beta = \mu, u = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \end{pmatrix}$$

The G matrix in the general mixed model describes the covariances among the random effects. In the animal model case we are looking at the covariances between relatives. The additive genetic covariance between relatives i and j is given as $2\Theta_{ij}\sigma_A^2$. Where Θ_{ij} is the coefficient of coancestry (see next section), and σ_A^2 is the additive genetic variance, or variance of breeding values, in the base population. This implies that under this animal model that $G = \sigma_A^2 A$ where A is known as the "Relationship Matrix" and has elements $A_{ij} = 2\Theta_{ij}$.

The covariance matrix for the residual errors, R in the general mixed model, is generally assumed to be $R = \sigma_E^2 I$ so each observation has the same variance and is uncorrelated with the other residual errors. This assumption can have many issues under the animal model, for example individuals being full sibs, or shared environmental effects, however these sorts of complications with residuals occur in almost any model, so for simplicities sake we will deal with the simple case where $R = \sigma_E^2 I$ and thus $R^{-1} = \sigma_E^{-2} I$.

Since in the animal model case $G^{-1} = \sigma_A^{-2} A^{-1}$ the Mixed Model Equations become the following for the animal model

$$\begin{pmatrix} X^T X & X^T Z \\ Z^T X & Z^T Z + \lambda A^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{u} \end{pmatrix} = \begin{pmatrix} X^T y \\ Z^T y \end{pmatrix}$$

where $\lambda = \sigma_E^2 / \sigma_A^2$. This lambda can also be expressed in terms of heritability, which frames lambda in terms of the coefficient of coancestry as defined in the next section. However, this still requires variance component estimation, so it only changes the calculation method of lambda slightly. In the simple case of the animal model we've been investigating the only fixed factor is μ, giving $\beta = \mu$ and $X = \mathbf{1}$, and in the case that each individual has only one observation, $Z = I$ with n individuals we further reduce this to

$$\begin{pmatrix} n & \mathbf{1}^T \\ \mathbf{1} & I + \lambda A^{-1} \end{pmatrix} \begin{pmatrix} \hat{\mu} \\ \hat{u} \end{pmatrix} = \begin{pmatrix} \Sigma^n y_i \\ y \end{pmatrix}$$

The other common model used is the Gametic Model, which is very similar to the Animal Model, however it concerns when breeding values of the parents are more important than the offspring. The additive genetic value of each offspring is expressed in terms of the parents' breeding values, with a_{si} being the breeding value for individual i's sire, and a_{di} the breeding value for individual i's dam. We then express the breeding value for i as

$$a_i = (\tfrac{a_{si}}{2} + \tfrac{a_{di}}{2}) + e_{ai}$$

which is the sum of the average of its parental value and a random deviation. This means we re-write the simple animal model investigated previously as

$$y_i = \mu + (\tfrac{a_{si}}{2} + \tfrac{a_{di}}{2}) + (e_{ai} + e_i)$$

2.7 Relationship Matrix

The relationship matrix A has in important role in the animal model, as it is present in the mixed model equations for the animal model, as well as an important piece of the variance structure. It is based on the genetic relationship between in individuals in the particular data set. This is often kept track of outside of the relationship matrix in a pedigree, or in common terms a "family tree". Unrelated individuals would have separate family tree's, while related ones would be connected in some way. This collection of tree's would make up the overall pedigree of the data set. There are methods for computing each of the individual elements of A, however this can become time consuming for large pedigrees. Henderson(1976) [18] showed that patterns exist that can be used to calculate the elements of A faster. He also showed that it is possible to obtain A^{-1} without having to compute A in the case of a non-inbred population.

We previously mentioned that the elements of A are $A_{ij} = 2\Theta_{ij}$, but we will now go into more detail as to what this value means. Look at the case where we have an offspring z, with parents x and y. Then Θ_{xy} is the coefficient of ancestry of x and y, which is the probability that one allele randomly drawn at a locus from x is identical by descent (IBD) from an allele drawn at random from y at the same locus. Two alleles being IBD implies that the two alleles are copies of the same ancestral allele. This

means that this coefficient is equivalent to Wright's(1922) [45] inbreeding coefficient of the offspring, written as f_z. There are several ways of calculating Θ_{xy} based on other measures of relatedness, however the way to calculate it directly can be shown with the calculation of Θ_{xx}, the coefficient of coancestry of an individual with itself. If we denote the two alleles carried by an individual as A_1 and A_2, randomly draw a allele, replace it, then randomly draw again, Θ_{xx} is the probability that the two alleles drawn are identical by decent. There are four ways alleles can be drawn, each with equal probability, with two of these ways being drawing the same allele twice, which is the situation we are looking for. Thus, if the two alleles are not identical Θ_{xx} is $\frac{1}{2}$. It is however possible that the individual is inbred. In this case the probability that A_1 is identical by descent to A_2 is f_x, so a general expression would be

$$\Theta_{xx} = \tfrac{1}{4}(1 + f_x + f_x + 1) = \tfrac{1}{2}(1 + f_x)$$

A more complex situation is one with calculating the coefficient between parent and offspring, call the parent of interest p and the offspring o. In the case where neither are inbred, then of the four ways single alleles can be drawn (one from p and one from o) only one gives a pair identical by descent. Thus $\Theta_{po} = \frac{1}{4}$. The case where the mother is inbred, the probability of both her alleles being identical by descent is f_p. This is the same as the probability that the parental allele inherited by the offspring is identical by decent to the allele not inherited. The probability of drawing this allele combination is $\frac{1}{4}$, giving a higher $\Theta_{po} = (1 + f_p)/4$. Complete inbreeding ($f_p = 1$) gives $\Theta_{po} = 1/2$. Finally if the two parents are related, o is now inbred with f_o, we now consider the probability of drawing a allele from the offspring coming from the non-parent of interest, the probability of this is $1/2$. Since f_o is equal to the probability that the maternal and parental alleles are identical by decent, the additional parent-offspring adds $\frac{f_o}{2}$, giving the general expression for the coefficent of ancestry between parent and offspring as

$$\Theta_{po} = \tfrac{1}{4}(1 + f_p + 2f_o)$$

Next is two individuals that share the same parents, known as full-sibs. Let s be the sire, d be the dam, and x and y the two offspring. If neither parent is inbred or related then there are two ways in which the same allele can be passed to both

offspring. If d passes the same allele to both, or if s passes the same allele to both. The probability that x and y receive the same allele from a particular parent is 1/2. Note that this is the coefficient of coancestry of a non-inbred parent with itself. Second, the probability of randomly drawing an allele from this parent from individual x is 1/2, and the same goes for individual y. Thus, the probability of drawing two alleles from a particular parent one from x and one from y, which are identical by decent is 1/8. The same process applies for the other parent, giving a total coefficient of ancestry $\Theta_{xy} = 1/4$. If we then go through the same process as previous, adding in the possibilities that s and/or d are inbred, we gain the following expression for the coefficient of ancestry for full sibs,

$$\Theta_{xy} = \tfrac{1}{8}(2 + f_d + f_s + 4\Theta_{sd})$$

These techniques can be extended to more complicated relationships between individuals. It is always calculated as as a sum of a series of two possible paths between x and y. One is leading from a single common ancestor to the two individuals, and the second goes through two ancestors that are related to each other. Neither can pass through the same ancestor more than once. The whole process is summarised as

$$\Theta_{xy} = \sum_i \Theta_{ii}(\tfrac{1}{2})^{n_i - 1} + \sum_j \sum_{j \neq k} \Theta_{jk}(\tfrac{1}{2})^{n_{jk} - 2}$$

This whole calculation process is outlined in detail in Genetics and Analysis of Quantitative Traits.[24]

It is quite clear the calculating each element in the relationship matrix can be quite tedious when large numbers of animals are involved. Henderson(1976) noted that certain patterns of allele flow through pedigrees can be used to expedite the construction of this matrix. As noted, $A_{ii} = 1 + f_i$ and $A_{ij} = 2\Theta_{ij} = 0$ if two individuals are unrelated. Following from these relationships Emik and Terrill(1949) [8] outlined rules that can be used to obtain elements of A for an arbitrary pedigree.

First, order the individuals so that patents precede their offspring, and allow the first b non-inbred, non-related individuals be known as the base population. The upper left $b \times b$ submatrix in the upper left of A is an identity matrix. We then expand this iteratively one row and column at a time until A is complete. If individual i has parents indexed by g and h then its diagonal element is

$$A_{ii} = 1 + f_i = 1 + \Theta_{gh} = 1 + \frac{A_{gh}}{2}$$

For a pair of individuals i and j, $j < i$

$$A_{ij} = A_{ji} = \Theta_{jg} + \Theta_{jh} = \frac{A_{jg}+A_{jh}}{2}$$

If a parent is not known, then assume it is non-inbred and non-related to any other measure individual (except known descendants), so if k is an unmeasured parent assume $A_{kk} = 1$ and $A_{ik} = 0$ for i where i is any individual except known descendants of k.

In many situations, both in nature and in breeding procedures only one parent is known with certainty. One way of dealing with this is to assume that the unknown parent is unrelated to any of the measured individuals in the base population. However, if all of the potential unknown parents are measured, ie, we have all the sires, but are unsure which sires produced which offspring, an average relationship matrix can be constructed, by assigning all of the potential parents an equal weight. Thus, if we had an individual i with a sire coming from k potential males, each of these is assumed to be the real sire with probability $1/k$, and so the entry for each sire in the ith row and column of A becomes $1/(2k)$. It is possible to use certain biological techniques to assign more accurate probabilities, however if the unknown parent can be reduced to a small number of potentials then the average relationship matrix is a powerful approach.

The above provides easier ways of calculating the relationship matrix A, but the problem still exist of the calculation of A^{-1} as it is A^{-1} that appears in the mixed model equations. This can be a complicated inversion if a large number of individuals are in the pedigree. This problem has caused a lot of attention to be focused on attempting to find short cuts, or easier ways of computing elements of A^{-1}. Henderson(1976) [18] showed that in a non-inbred population the inverse of A can be obtained without having to compute A itself. For n individuals, order n operations are required to calculate A^{-1} by Hendersons method, which n^2 and n^3 operations are required to calculate A and then A^{-1} with normal methods. Henderson's main method is that the relationship matrix can be expressed as

$$A = TDT^T$$

and that its inverse is

$$A^{-1} = (T^{-1})^T D^{-1} T^{-1}$$

D is a diagonal matrix, the elements of which are are proportional to variances associated with segregational sampling conditional on the parents, which are easily acquired in a non-inbred population. In this case D_{ii} is 0.5, 0.75, or 1.0 when both, one, or none of individual i's parents are included in A. D is diagonal, so its inverse is diagonal with its elements being the reciprocal of the elements in D. T is a lower triangular matrix, whose elements trace the flow of genes through the sample. The diagonal elements are all equal to one, while the elements in the jth row in the column below the ith diagonal are defined as the fraction of the genes of individual j that that are expected in individual i. In nonrelated individuals this value would be 0, and in non-inbred populations the elements would be 1/2, 1/4, 1/8 for first, second, and third degree relatives, and continuing in this pattern. T^{-1} is also lower triangular, with ones on the diagonal, and below the diagonal in the jth row all elements are zero, except those corresponding to the column of j's known parents, which are equal to -0.5.

These rules allow for a much faster computation of A^{-1}, eliminating the need to invert A normally. Methods proposed by Quaas(1976) [34], Tier(1990) [41] and Mrode(2014) [31] extend this concept to allow for inbreeding. This work on reducing the difficultly in calculation of A^{-1} allows for much faster solutions to the mixed model equations when dealing with the animal models, and allows this BLUP methodology to be used in very large and complicated data sets in animal breeding, where the pedigrees can contain numerous entries, eg in dairy cattle where the numbers can reach hundreds of thousands or millions.

2.7.1 Examples

Mentioned above was that these relationship matrices can be very large very quickly. This is due to the nature of their construction in that they relate every individual in the data set to every other individual in the set. This will create an nXn matrix in a data set with n individuals. Thus when data sets get large, as they often do in the animal breeding context, very large matrices are created. As a small example of how

Table 2.1: Simple Example Pedigree

Animal	Sire	Dam
1	-	-
2	-	-
3	-	-
4	1	2
5	3	2

Table 2.2: Example Pedigree - Slight Inbreeding

Animal	Sire	Dam
1	-	-
2	-	-
3	1	-
4	1	2
5	3	4
6	1	4
7	5	6

these are constructed, if we take 5 individuals related according to the pedigree given in Table 2.1, we obtain the relationship matrix in equation 2.7.1. Note that in the pedigree animals listed as a "-" are unknown, and in these examples are considered to be non-inbred and unrelated.

$$A = \begin{pmatrix} 1 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 1 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 & 1 & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{4} & 1 \end{pmatrix} \tag{2.7.1}$$

A slightly more complicated example, including some inbreeding and more individuals as detailed in the pedigree in table 2.2, provides the relationship matrix in equation 2.7.2.

$$A = \begin{pmatrix} 1 & 0 & 0.5 & 0.5 & 0.5 & 0.75 & 0.625 \\ 0 & 1 & 0 & 0.5 & 0.25 & 0.25 & 0.25 \\ 0.5 & 0 & 1 & 0.25 & 0.625 & 0.375 & 0.5 \\ 0.5 & 0.5 & 0.25 & 1 & 0.625 & 0.75 & 0.6875 \\ 0.5 & 0.25 & 0.625 & 0.625 & 1.125 & 0.5625 & 0.84375 \\ 0.75 & 0.25 & 0.375 & 0.75 & 0.5625 & 1.25 & 0.90625 \\ 0.625 & 0.25 & 0.5 & 0.6875 & 0.84375 & 0.90625 & 1.28125 \end{pmatrix} \tag{2.7.2}$$

2.8 Joint Estimation

The BLUP model as described so far can be further extended in several ways. The first that will be mentioned here is extending the model to situations where two or more vectors of random effects are of interest. Expressing these two vectors as u_1 and u_2, which are uncorrelated with each other, the mixed model then becomes

$$Y = X\beta + Z_1 u_1 + Z_2 u_2 + e$$

The vectors of random effects can have different dimensions (say q_1 and q_2), with n individuals in y we have Z_1 as $n \times q_1$ and Z_2 is $n \times q_2$, using the same notation as previous but with G_i being the $q_i \times q_i$ covariance matrix for u_i the MME's then are

$$\begin{pmatrix} X^T R^{-1} X & X^T R^{-1} Z_1 & X^T R^{-1} Z_2 \\ Z_1^T R^{-1} X & Z - 1^T R^{-1} Z_1 + G_1^{-1} & Z_1^T R^{-1} Z_2 \\ Z_2^T R^{-1} X & Z_2^T R^{-1} Z_1 & Z_2^T R^{-1} Z_2 + G_2^{-1} \end{pmatrix}$$

$$\begin{pmatrix} \hat{\beta} \\ \hat{u_1} \\ \hat{u_2} \end{pmatrix} = \begin{pmatrix} X^T R^{-1} y \\ Z_1^T R^{-1} y \\ Z_2^T R^{-1} y \end{pmatrix}$$

This can be extended to include additional uncorrelated vectors of random effects if desired, following the same pattern. An example of this is the Maternal Effects model, where the phenotype of an individual is modelled on both genetic and environmental components of maternal effects, as well as an individuals direct genetic contributions. In this case, there end up being three vectors of random effects, one

for each component mentioned. For more information on this model in particular, see Quaas and Pollak(1981) [35].

Another possible extension is that of multivariate BLUP. In the particular case of breeding value estimation with multiple traits, it is possible to preform univariate BLUP on each of the individual traits, however this is inefficient. Often it is the case that traits can be correlated, and one can provide information about others. Dealing with BLUP in a multivariate context takes this information into account using conditional expectations, but does require accurate estimates of genetic and environmental covariances among the traits and has high computational demands. For further insight into multivariate BLUP methods, see Henderson and Quaas(1976) [19].

2.9 Variance-Component Estimation

When dealing with BLUP's in the genetic context, variance components are often estimated using Maximum Likelihood estimation (ML), or Restricted maximum likelihood (REML), rather than ANOVA. This is due to a few factors, namely that ANOVA generally requires sample sizes be well balanced. However this is often not the case when dealing with quantitative genetics, as individuals are often lost, or group sizes aren't similar to begin with. The need for this isn't present in ML or REML. The other reason ML and REML is preferred is related to the structure of the pedigree, since individuals are often strongly related to one another, it is not at easy to analyze them jointly with ANOVA as it is other methods.

We will now investigate the particular use of ML and REML in the context of the general mixed model that we've been investigating. In general in the genetic context REML is used far more frequently than ML, but it is useful to start with ML, as REML can be expressed as a linear transform of a ML problem.

Start with the general mixed model as before, $y = X\beta + Zu + e$ where $u \sim \text{MVN}(0, G)$ and $e \sim \text{MVN}(0, R)$, this implies that $y \sim \text{MVN}(X\beta, V)$ where $V = ZGZ^T + R$. We then obtain the probability density of y as

$$p(y|X\beta, V) = (2\pi)^{-\frac{n}{2}}|V|^{-\frac{1}{2}}\exp\left[-\tfrac{1}{2}(y - X\beta)^T V^{-1}(y - X\beta)\right]$$

We then take the natural log of the above to obtain the log-likelihood of β and V given the observed data

$$L(\beta, V|X, y) = -\tfrac{n}{2}\ln(2\pi) - \tfrac{1}{2}\ln|V| - \tfrac{1}{2}(v - X\beta)^T V^{-1}(y - X\beta)$$

This is the log-likelihood for the general mixed model as above, however for the following discussion we will consider the genetic case where $u = a$ is a vector of additive genetic values (breeding values). We then try to estimate the variance components that are in G and R, namely $G = \sigma_A^2 A$ where A is the relationship matrix, and $R = \sigma_E^2 I$, as well as adding the possibility of extending the model as spoken of before to the following

$$y = X\beta + \sum_{i=1}^{m} Z_i u_i + e$$

here we have m vectors of random effects, all assumed to be uncorrelated with $u_i \sim \text{MVN}(0, \sigma_i^2 B_i)$, where B_i is a matrix of known constants. The log-likelihood is still given as above, but now

$$V = \sum_{i=1}^{m} \sigma_i^2 Z_i B_i Z_i^T + \sigma_E^2 I$$

Using the results for matrix derivatives taking the derivative of the log-likelihood with respect to β is

$$\tfrac{(\beta, V|X, y)}{\partial \beta} = X^T V^{-1}(y - X\beta)$$

before we do the variance components if we rewrite in terms of deviations as follows it becomes easier to see the bias in the ML estimates more easily.

$$(y - X\beta)^T V^{-1}(y - X\beta) = (y - X\hat{\beta})^T V^{-1}(y - X\hat{\beta}) + (\hat{\beta} - \beta)^T X^T V^{-1} X(\hat{\beta} - \beta)$$

Take the above and substitute it into the log-likelihood then taking the derivative with respect to the variance components gives

$$\begin{aligned}\tfrac{\partial L(\beta, V|X, y)}{\partial \sigma_i^2} = &-\tfrac{1}{2}\text{tr}(V^{-1}V_i) + \tfrac{1}{2}(y - X\hat{\beta})^T V^{-1} V_i V^{-1}(y - X\hat{\beta}) \\ &+\tfrac{1}{2}(\hat{\beta} - \beta)^T X^T V^{-1} V_i V^{-1} X(\hat{\beta} - \beta)\end{aligned}$$

where

$$V_i = \frac{\partial V}{\partial \sigma_i^2} = \begin{cases} I & \text{when } \sigma_i^2 = \sigma_E^2 \\ Z_i B_i Z_i^T & \text{otherwise} \end{cases}$$

We then take the previous derivatives and set them equal to zero and solve we obtain the ML estimate of the fixed effects, $\hat{\beta} = (X^T\hat{V}^{-1}X)^{-1}X^T\hat{V}^{-1}y$ which is the same as the BLUE obtained earlier for the fixed effects. More interesting is setting $\hat{\beta} = \beta$ in the derivative with respect to the variance components and rearranging as follows to obtain the ML estimate of the variance components.

$$\text{tr}(\hat{V}^{-1}V_i) = (y - X\hat{\beta})^T\hat{V}^{-1}V_i\hat{V}^{-1}(y - X\hat{\beta}) \tag{2.9.1}$$

Simplifying the above using the matrix

$$P = V^{-1} - V^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}$$

we obtain

$$\text{tr}(\hat{V}^{-1}V_i) = y^T\hat{P}V_i\hat{P}y$$

where we use $\hat{P}$ as a reminder that P is a function of V, which we are trying to estimate. So with the m random effects, the set of $m + 1$ ML equations for the variances of random effects are

$$\text{tr}(\hat{V}^{-1}) = y^T\hat{P}\hat{P}y \text{ for } \sigma_E^2$$
$$\text{tr}(\hat{V}^{-1}Z_iB_iZ_i^T) = y^T\hat{P}Z_iB_iZ_i^T\hat{P}y \text{ for } \sigma_i^2,\ 1 \leq i \leq m$$

with $\hat{P}$ using

$$\hat{V} = \sum_{i=1}^{m} \hat{\sigma}_i^2 Z_iB_iZ_i^T + \hat{\sigma}_E^2 I \tag{2.9.2}$$

Note that unfortunately $\hat{\beta}$ is dependant on $\hat{V}$ which is itself contains variance components that we wish to estimate. The solutions also contain $\hat{V}^{-1}$, which means that they are non-linear functions of these variance components. This gives the result that there is no simple solution. However the standard errors of the ML estimates can be obtained from the Fisher information matrix as normal in the theory of Maximum Likelihood, these are given as

$$\sigma(\beta_i\beta_j) = (X^TV^{-1}X)^{-1}_{ij} \text{ and } \sigma(\sigma_i^2\sigma_j^2) = (S^{-1})_{ij} \text{ where}$$
$$S_{ij} = \tfrac{1}{2}\text{tr}(V^{-1}V_iV^{-1}V_j) \text{ where } V_i \text{ is given as above.}$$

The ML estimates for fixed and random effects are uncorrelated. That is, $\sigma(\beta_i,\sigma_j) = 0$.

REML on the other hand, is based on a linear transformation of the observations y, that removes the fixed effects from the model. To do this, say we have a matrix K associated with the design matrix X such that $KX = 0$. If we apply this transformation to the mixed model, we obtain the following

$$y^* = Ky = K(X\beta + Za + e) = KZa + Ke$$

In this case, y^* is equivalent to the residual deviations from the estimated fixed effects. That is $y_i^* = y_i - \bar{y}$. Since REML is just a linear transform of ML estimates, REML estimates of variance components are the same as ML estimates. Therefore, we can use the ML solutions above but with the following transformations

$$Ky \text{ for y, } KX = 0 \text{ for } X, \; KZ \text{ for } Z, \; KVK^T \text{ for } V$$

It seems like we have added an extra complication in that we now need to calculate the K matrix as well, however Searle et al.(1992) [37] showed that K satisfies

$$P = K^T(KVK^T)^{-1}K$$

Then note that

$$(y^*)^T(V^*)^{-1}y^* = (y^TK^T)(KVK^T)^{-1}(Ky) = y^TPy$$

Substituting this into equation (2.9.1), and performing some rearrangement the ML equations yield the REML estimators as

$$\text{tr}(\hat{P}) = y^T\hat{P}\hat{P}y \text{ for } \sigma_E^2$$
$$\text{tr}(\hat{P}ZAZ^T) = y^T\hat{P}ZAZ^T\hat{P}y \text{ for } \sigma_A^2$$

REML does not give estimates of β, since we remove the fixed effects by setting $\beta^* = 0$.

This transformation $y^* = KY$ only depends on the design matrix, and this approach still holds in the m uncorrelated random vectors case as before, with the equation then being

$$y^* = \sum_{i=1}^{m} KZ_i u_i + Ke$$

and the $m+1$ REML equations for the variance components become

$$\text{tr}(\hat{P}) = y^T \hat{P}\hat{P}y \text{ for } \sigma_E^2$$
$$\text{tr}(\hat{P}Z_i B_i Z Z_i^T) = y^T \hat{P} Z_i B_i Z_i^T \hat{P} y \text{ for } \sigma_i^2,\ 1 \leq i \leq m$$

where $\hat{P}$ is now a function of $\hat{V} = \sum \hat{\sigma}_i^2 Z_i B_i Z_i^T + \hat{\sigma}_E^2 I$.

In REML, the information matrix has no elements corresponding to the fixed effects, only the variance component estimates. This means that the fisher information matrix F is $F = S$ where S is

$$S_{ij} = \tfrac{1}{2}\text{tr}(PV_iPV_j)$$

where V is given as before (2.9.2). Estimates of the sampling variance and covariance of the variance-component estimates are obtained by the inversion of S.

As above, REML methods can be extended to the multivariate case. Similar log-likelihoods are constructed and solved. For details on this see, Meyer(1985) [26], Schaeffer(1986) [36], Taylor et al(1985) [39], Jensen and Mao(1988) [20], and Thompson and Hill(1990) [40].

Mentioned above was the fact that the ML and REML equations are not straightforward to solve, as they are non-linear. Only in specific cases, namely completely balanced designs, are closed solutions able to be found. It is possible to find solutions using grid searching, however this can be computationally difficult when the number of elements of β increases, as each element will increase the dimensionality of the search. REML reduces this, as it eliminates β entirely, but the REML likelihood function is much harder to compute. Thus, numerical, iterative solving methods are generally used, such as the Newton-Raphson algorithm and the EM algorithm. These methods can still be computationally intense when we have large pedigrees, as every step will require large matrix inversions. Detailed reviews of different methods to solve these equations can be found in Meyer(1989b) [28], Harville and Calllanan(1990) [14], and Searle et al.(1992) [37].

There also exist several pieces of software designed to solve REML equations for modelling purposes. The common commercial option is ASREML by VSN international. This is a commercially available product in common use in the biosciences

for analysis of mixed models, designed for use with genetic based data sets, and includes a built in R link to make programming easier for users who prefer to use the R programming environment. A second common choice is WOMBAT, written and maintained by Karin Meyer [29]. It is specifically designed to fit linear mixed models using REML techniques to estimate covariance components. Once again, it can be applied to many different problems and data sets, but it's main purpose is to deal with genetic models. Major common statistical analysis softwares have packages to deal with REML, such as R's nlme, and lme4; SAS, STATA, and SPSS also have ways of performing REML analysis. Other stand alone software can deal with REML effectively, however ASREML and WOMBAT were designed for genetic data in particular, hence the special mention.

Finally, an important result is that Cressie and Lahiri(1993) [6] give conditions under which it is possible to show that REML estimates are asymptotically normal, with mean zero, and a variance equal to the inverse of the restricted information matrix. They do this using a slightly different notation then we do, taking the random effects and incorporating them into the error structure, such as we showed Goldberger did previously. The notation they use is $Y \sim N(X\beta, \sum(\theta))$. They do so by first deriving the REML estimates as above, however in a more general context rather than the particular case of genetics like we use. They then use some results from Sweeting(1980) [38] that were developed for regular maximum likelihood estimation and its asymptotic properties to show that REML estimators are asymptotically normal under certain conditions. These conditions are reasonable, mostly requiring smoothness of the variance matrix $\sum(\theta)$, requiring that it is twice differentiable. In our notation this would be a requirement of the variance matrix V. This is an important result, as we now know that REML estimation can be equally as accurate as ML estimation, knowing its asymptotic properties.

Chapter 3

Social Interaction and Simulation Results

Mentioned previously was that the model proposed by Bijma an Muir contains factors to account for the "social effect" that fish in an aquaculture setting experience. It was discussed how important these factors can be in this setting, especially when compared to other animal breeding programs. This model is the focus of our study, and is used as the tool to analyse simulated data in order to determine its estimation effectiveness. Here we will explain the basics of what are known as Indirect Genetic Effects, and Direct Genetic effects, as these are the two main components of the proposed model. We then will go into more detail about this model and its particulars, as well as starting the simulation and exploring the results obtained.

3.1 IGE's and DGE's

Social interactions in general, such as competition or cooperation have been shown to be important to natural selection (Darwin(1859) [7], Hamilton(1964) [13], Wilson(1975) [43], Frank(1998) [10], Keller(1999) [23], Clutton-Brock(2002) [4]), however most research focus in this area has been on fitness-effects on individuals, and success of populations. Evidence suggests, that social interactions between individuals in either natural populations, or when in captivity, can cause individual phenotypes to depend on the genetic make up of other individuals (Wolf et al.(1998) [44], McGlothlin and Brodie(2009) [25]). These sorts of effects are knows as Indirect Genetic Effects, or IGE's, namely a heritable trait of one individual, that has an effect on a trait value of another individual. An example is maternal genetic effect of a mother on preweaning growth weight of her offspring in mammals (Willham(1963) [42], Mousseau and Fox(1998) [30]). IGE's may have very important effects on response to natural or artifical seletion, in some cases these can be very drastic effects(Griffing(1967))'[12], and thus knowing how they interact with and relate to Direct Genetic Effects, or DGE's, is useful for understanding this response to selection.

Mixed models can be used to estimate the magnitude of IGE's, and their contribution to heritable variance within populations. These models allow to estimate IGE's without the need to observe the interactions, or even knowledge of what in particular they are. This is due to the splitting of fixed and random effects in mixed models, allowing us to determine the magnitude of the effects on heritable variance from IGE's and DGE's, and then compare them to each other. In general, knowledge about this relationship is still limited, and more needs to be done to understand IGE's, and how they contribute to natural selection and genetic improvement in populations for the breeding of animals or plants.

Bijma(2010) [2] suggests using a variance component model, with the genetic variance in trait value split up into a direct component due to individual genotype, and an indirect component due to the genotypes of other individuals in the social group (Willham(1963) [42], Griffing(1967) [12]). This model is as follows.

Consider a population with groups of n_w members, with interactions occurring between the members of each group. The trait value of individual i is expressed as the sum of a direct effect due to that individual, $P_{D,i}$, and the sum of the indirect effects $P_{S,j}$ of each of the other members of the group.

$$P_i = P_{D,i} + \sum_{j=1}^{n_w-1} P_{S,j}$$

Thus, i is the focal individual, j is one of its group mates, D is direct effects, S indirect effects and the sum is taken over the $n_w - 1$ fellow group members of individual i. He then restructures this into an additive genetic component (heritable) A, and a non-heritable component E.(Griffing(1967) [12]).

$$P_i = A_{D,i} + E_{D,i} + \sum_{j\neq i}^{n_W-1} A_{S,j} + \sum_{j\neq i}^{n_W-1} E_{S,j}$$

Here, $A_{D,i}$ is the DGE of the individual in focus, $E_{D,i}$ is the non-heritable direct effect for i, $A_{S,j}$ is the IGE for group member j, and $E_{D,j}$ is the non-heritable indirect effect from j.

On a larger scale, at the population level, these IGE's and DGE's are often measured by their covariances, as we are often interested in the population as a whole, rather than particular individuals. Thus we are interested in the variance of DGE

$\sigma^2_{A_D}$, of IGE $\sigma^2_{A_S}$ and their covariance $\sigma_{A_{DS}}$. In particular for Bijma, the total heritable variance due to the joint effect of DGE's and IGE's is of interest, and the total impact of an individual's genes on the mean trait value of a population, (size n), is given by the individual's total breeding value (TBV) (Bijma et al.(2007) [3]).

$$TBV_i = A_{D,j} + (n-1)A_{S,j} \tag{3.1.1}$$

Note that in equation 3.1.1, TBV is a heritable property of an individual, as it is a more general form of the classical breeding value, and thus it is the relevant property for selection in traits that are affected by IGE's.

Then the total heritable variance in the trait due to both IGE's and DGE's is equal to the variance in TBV's among individuals (Bijma et al.(2007) [3]).

$$\sigma^2_{TBV} = \sigma^2_{A_D} + 2(n_W - 1)\sigma_{A_{DS}} + (n_W - 1)^2\sigma^2_{A_S}$$

Interpretation of σ^2 is often done by expressing heritable variance relative to phenotype variance. Heritability measures heritable variance in relation to phenotypic variance, $h^2 = \sigma^2_A/\sigma^2_P$. The same relationship can apply in the case of IGE's. In this case we can define the ratio of total heritable variance over phenotypic variance (Bergsma(2008) [1]).

$$\tau^2 = \frac{\sigma^2_{TBV}}{\sigma^2_P}$$

3.2 Simulation Details

Investigating how this model that includes IGE's and DGE's behaves under various conditions is important so that interpretation of experimental results can have more meaning. Thus, as simulation was run based on this model in order to determine accuracy of prediction. There are five important parameters involved in the parameter space, those being, σ^2_{dd}, σ^2_{ss} , σ^2_{sd}, σ^2_c, and σ^2_e. Namely, the variance of the direct genetic effect, the variance of the indirect genetic effect (or social effect for simplicity), the covariance of these two, the variance of the fixed "cage" effect, and the error. This will be the notation used going forward, but these are the same values as mentioned

in the previous section, which used Bijma's notation. What we want to do is see how well the model can predict these terms under differing initial conditions for them. In particular we wish to focus on the direct, and social effects, as well as their covariance, as these are the interesting terms. The cage effect is a fixed effect, and error is a random process, and thus they do not particularly help us in determining how the IGE's and DGE's interact with each other. However controlling for them in our modelling will ideally allow us to obtain more accurate results.

The simulation was set up to attempt to mirror theoretical optimums Bijma[2] suggested while still being reasonable for data generation. His conclusion was that that 2 families per cage would be optimal for accuracy. The data was structured using a few parameters that allow generation of the necessary data structure, the most important of which is the pedigree, or the relationship between each individual fish. For the purposes of simplifying the analysis, we generate all of the fish as full-sibs, meaning offspring from a family share the same two parents. This, plus assuming that none of the offspring are inbred creates the simplest pedigree for analysis. For each run of the simulation data for each individual fish was generated, using 30 dams, with one sire per dam, creating 30 individual families. The number of groups or cages was determined by assigning 2 families per cage, with 5 individuals from each family in these cages. In order to attempt to reduce the number of cages while still exposing each family to a large amount of the other families, the family paring per cage were assigned in a cyclical manner. Finally, a second block following the same structure as this was also generated to create some replication.

Fixed, "true" values were then set for the five parameters in order to generate data based on the experimental design as above. The generated data was then analysed using Karin Meyers' WOMBAT software [29] for analysing mixed models using BLUP and the mixed model proposed by Bijma. This allowed an estimation of the five parameters such that we can compare them to the initial set values, and we can determine accuracy of prediction. In particular, it is interesting to determine how different levels of different parameters interact with each other. Thus, multiple simulations with different initial values is valuable. This leads naturally to a factorial design analysed with ANOVA.

This was done by selecting high and low initial values for each of the parameters

and creating a 2^5 factorial design. Each of these 32 designs were then simulated as above with 500 repetitions for each parameter set, generating new data using the same generation scheme described above each time, then analysed through ANOVA in two different ways. The first was done by taking the squared error between the initial "true" value of a particular parameter, and the estimated value for that parameter as the response, with the levels of all of the parameters used as the factors. This gives 16000 data points with which to infer from if the selected parameters estimated error has any relationship to the initial conditions of all five parameters. The second way was just using the estimated value itself rather than the error, as patterns could perhaps be perceived that way as well. This was done for 3 of the interesting parameters, the Direct Variance, the Social Variance and the Covariance between the two. The error and cage effects were left out as we are interested in DGE's and IGE's, rather than fixed and error effects. However these effects were included in the modelling for accuracy purposes. This gives a total of 6 base ANOVA analyses, two for each of the parameters. More analysis was done to each parameter estimation after these initial ANOVA's to see if better results could be obtained. Details for each individual parameter are as follows in Table 3.1.

Table 3.1: Design Table for the 2^5 Factorial Design

Design	SS	SD	DD	E	C
1	0.001	0.0	0.1	0.1	0.1
2	0.009	0.0	0.1	0.1	0.1
3	0.001	0.3	0.1	0.1	0.1
4	0.009	0.3	0.1	0.1	0.1
5	0.001	0.0	0.9	0.1	0.1
6	0.009	0.0	0.9	0.1	0.1
7	0.001	0.3	0.9	0.1	0.1
8	0.009	0.3	0.9	0.1	0.1
9	0.001	0.0	0.1	0.4	0.1
10	0.009	0.0	0.1	0.4	0.1
11	0.001	0.3	0.1	0.4	0.1
12	0.009	0.3	0.1	0.4	0.1
13	0.001	0.0	0.9	0.4	0.1
14	0.009	0.0	0.9	0.4	0.1
15	0.001	0.3	0.9	0.4	0.1
16	0.009	0.3	0.9	0.4	0.1
17	0.001	0.0	0.1	0.1	0.4
18	0.009	0.0	0.1	0.1	0.4
19	0.001	0.3	0.1	0.1	0.4
20	0.009	0.3	0.1	0.1	0.4
21	0.001	0.0	0.9	0.1	0.4
22	0.009	0.0	0.9	0.1	0.4
23	0.001	0.3	0.9	0.1	0.4
24	0.009	0.3	0.9	0.1	0.4
25	0.001	0.0	0.1	0.4	0.4
26	0.009	0.0	0.1	0.4	0.4
27	0.001	0.3	0.1	0.4	0.4
28	0.009	0.3	0.1	0.4	0.4
29	0.001	0.0	0.9	0.4	0.4
30	0.009	0.0	0.9	0.4	0.4
31	0.001	0.3	0.9	0.4	0.4
32	0.009	0.3	0.9	0.4	0.4

3.3 ANOVA and Other Results

3.3.1 Direct Variance

Figure 3.16 at the end of the chapter shows a figure containing boxplots of the direct genetic variance estimates for the 32 designs used. The green line on the plot is a marker to show the true value for the variable's low value, while the red is representative of the high value. In general it seems that in most of the designs the estimates were fairly accurate, with Designs 3, 4, and 13 being the least accurate. The former two looking as if they are close to the high value, when they should be towards the low value, and design 13 at the low, while it should be near the high value. There are a number of designs with large outliers, but it is noteworthy that there are patterns of designs with almost none. These are the designs that had a covariance in the initial parameter set, rather than no covariance. These outliers are likely due to

Table 3.2: Direct Error ANOVA After Selection

Effect	Df	Sum Sq	Mean Sq	F value	Pr(> F)
SS	1	0	0.06	0.143	0.7057
SD	1	200	200.20	441.165	< 2e-16
DD	1	0	0.02	0.043	0.8359
E	1	0	0.15	0.339	0.5603
C	1	12	12.28	27.067	1.99e-07
SS:DD	1	7	7.36	16.213	5.69e-05
SD:DD	1	53	52.94	116.668	< 2e-16
SD:E	1	43	43.47	95.785	< 2e-16
DD:E	1	71	71.40	157.331	< 2e-16
SD:C	1	13	13.13	28.935	7.59e-08
DD:C	1	17	16.71	36.825	1.32e-09
E:C	1	17	16.69	36.787	1.35e-09
SS:SD:DD	1	2	2.25	4.950	0.0261
SS:DD:C	1	2	1.95	4.288	0.0384
SD:DD:C	1	15	14.75	32.513	1.21e-08
SD:E:C	1	15	14.99	33.037	9.21e-09
DD:E:C	1	25	24.73	54.489	1.64e-13
SS:SD:DD:C	1	2	1.94	4.281	0.0386
SD:DD:E:C	1	2	2.43	5.363	0.0206
Residuals	15970	7247	0.45		

REML failing to converge, but it is interesting that the estimation variability seems to collapse in the case where covariance is present. However this behaviour shows itself in boxplots for the social variance as well, and may be due to the nature of the data generation.

There were several ANOVA's performed, and starting with the Direct Genetic Effect variance ANOVA we see the results in Table 3.2. These results are shown after model selection procedures. There is a lot of significance, particularly in many of the second and third order interaction terms. The reason that these significances appear is likely due to the minor significance of the two 4^{th} order interaction terms. In general, all of the ANOVA's performed should have fairly significant results, as we have a very large sample size, giving a large residual degrees of freedom and decreasing the size of the p-values.

However, if we look at the QQ plot for the ANOVA in Figure 3.1, we can see

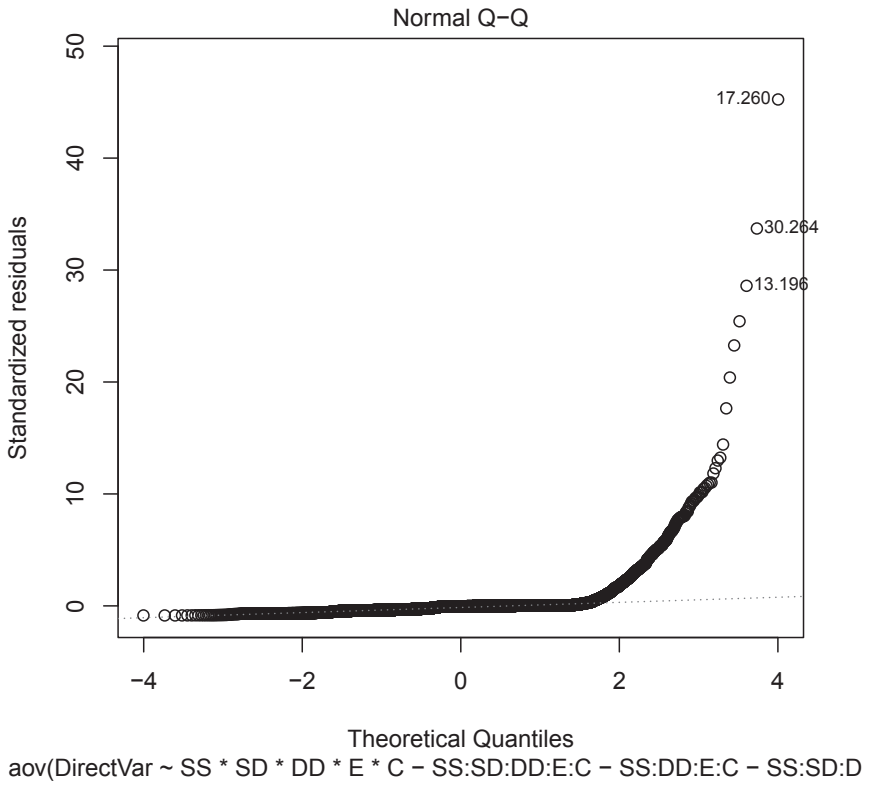

Figure 3.1: QQ Plot for Direct Error ANOVA After Selection

issues, namely with an incredibly long upper tail that seems to imply deviation from normality. This would lead to attempting to explore other options for analysis, namely attempting transforms to remove the tailed nature. A log transform of the data was attempted and then analysed, yielding the ANOVA results in Table 3.3. We see essentially the same results as above, with significant high order interactions, albeit more significant. The same problem arises in the QQ plot in Figure 3.2 as well, with large tails that still seem to deviate too much.

Table 3.3: Log Transform of Direct Error ANOVA After Selection

Effect	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SS	1	50	50	12.114	0.000502
SD	1	44085	44085	10649.845	< 2e-16
DD	1	1	1	0.138	0.710101
E	1	21731	21731	5249.583	< 2e-16
C	1	23947	23947	5784.992	< 2e-16
SS:DD	1	244	244	58.980	1.68e-14
SD:DD	1	23550	23550	5689.182	< 2e-16
SD:E	1	13919	13919	3362.553	< 2e-16
DD:E	1	19147	19147	4625.472	< 2e-16
SD:C	1	18921	18921	4570.741	< 2e-16
DD:C	1	949	949	229.259	< 2e-16
E:C	1	34546	34546	8345.499	< 2e-16
SS:SD:DD	1	238	238	57.531	3.51e-14
SS:DD:C	1	85	85	20.461	6.13e-06
SD:DD:C	1	1608	1608	388.554	< 2e-16
SD:E:C	1	37712	37712	9110.242	< 2e-16
DD:E:C	1	61	61	14.642	0.000130
SS:SD:DD:C	1	84	84	20.205	7.01e-06
SD:DD:E:C	1	6383	6383	1542.021	< 2e-16
Residuals	15970	66108	4		

Since we seem to be having residual issues in both the transformed and untransformed ANOVA's it is perhaps worth investigating them further in order to determine if a subset of these residuals is causing issues. The histogram in Figure 3.3 is of the residuals for the ANOVA results contained in Table 3.2. It seems to show that the large majority of the residuals are in a small range with only a few being large. If we took the range $(-2, 2)$, we obtain approximately 300 residuals that fall outside this range. We can then bin the residuals, setting their value to 1 if they are inside of the range, and 0 otherwise. A Logit regression was then performed on these binned values versus the levels of the design variables to see if particular designs were the

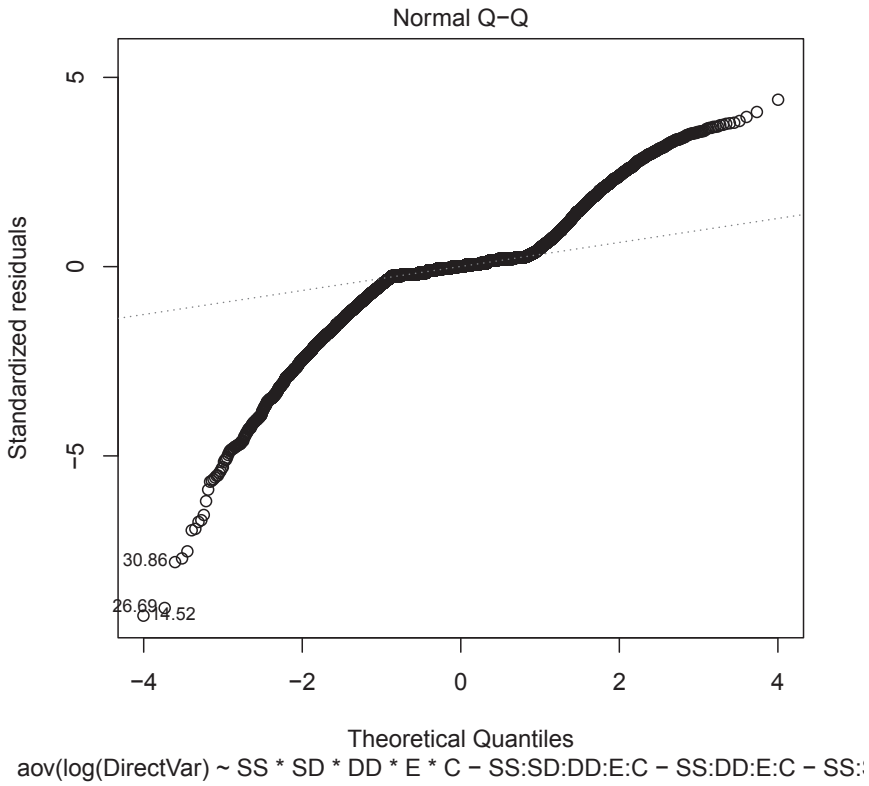

Figure 3.2: QQ Plot for Log Transform of Direct Error ANOVA After Selection

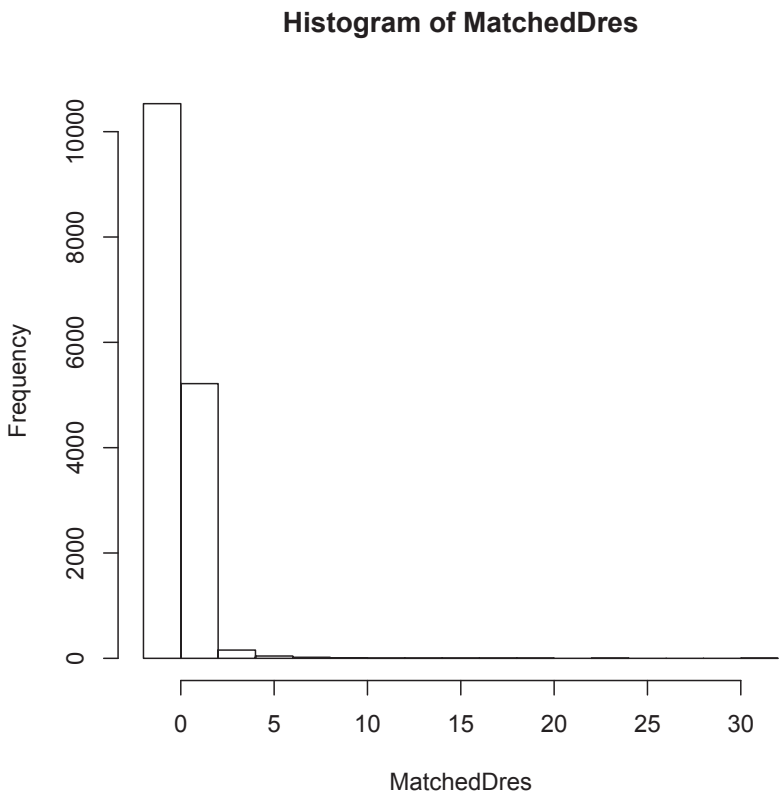

Figure 3.3: Histogram of Residuals for the Non-Transformed Direct Error ANOVA

ones causing the larger residuals, as this would be useful in both determining if the model breaks down under particular conditions, or if there are perhaps better ways to use this data for prediction.

Table 3.4 is the results table for the Logit regression performed on the binned residuals. There are some highly significant terms, however we can see the same issue with the QQ plot in Figure 3.4, where a large tail deviates particularly strongly, which doesn't seem to indicate any particular pattern to the residuals or a potential solution to the residual issues.

Table 3.4: Direct Error Logit Regression on Binned Residuals

Effect	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.5019	0.3635	9.635	< 2e-16
SS	-139.6152	35.3622	-3.948	7.88e-05
SD	10.7489	1.0774	9.977	< 2e-16
DD	-0.5880	0.5514	-1.066	0.286267
E	2.1552	0.9871	2.183	0.029003
C	3.0128	1.0399	2.897	0.003766
SS:DD	365.6609	75.2806	4.857	1.19e-06
DD:E	5.6722	1.7606	3.222	0.001274
DD:C	-3.3128	1.1319	-2.927	0.003425
E:C	-14.4671	3.7925	-3.815	0.000136
SS:DD:E	-1341.7519	239.9798	-5.591	2.26e-08
SS:E:C	1125.8528	341.9474	3.292	0.000993

It may be worth attempting to remove some of the problem residuals. While removing information is not often a good solution, in our case we are simulating data, and some of the bad residuals may be cases of the model not converging well for that particular run. Removing some of the bad residuals could give insight to weather or not this was the case. Table 3.5 is the ANOVA table for the Direct Effect variance error estimation, however with all of the residuals outside of the $(-2, 2)$ range removed.

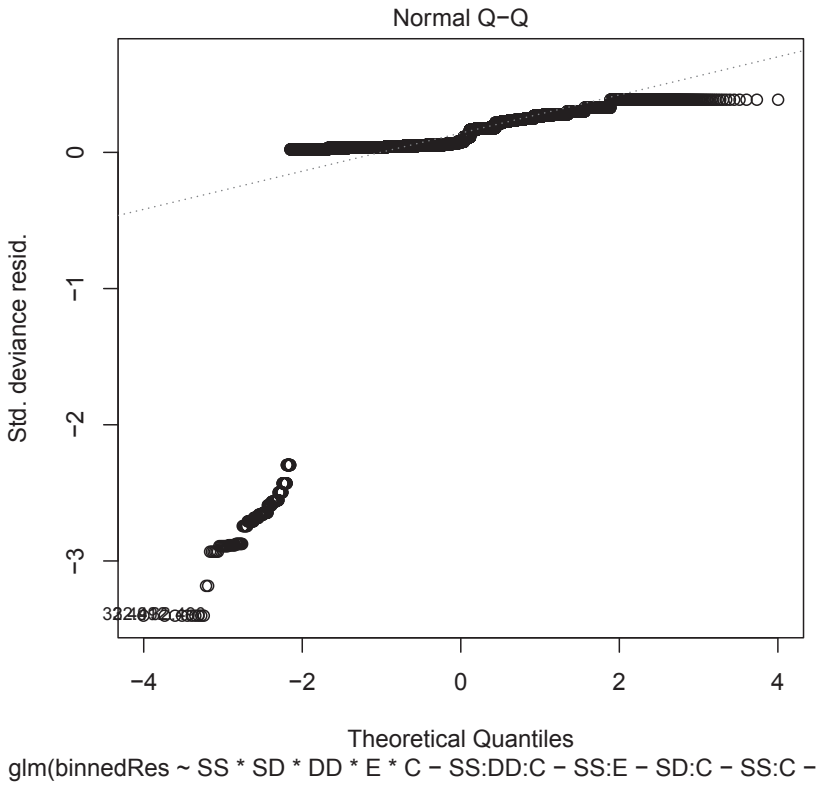

Figure 3.4: ANOVA Table for LOGIT Regression of Binned Residuals

Table 3.5: Direct Error ANOVA After Selection With Removed Residuals

Effect	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SS	1	0	0.27	0.611	0.434352
SD	1	191	190.84	436.031	< 2e-16
DD	1	0	0.08	0.184	0.668123
E	1	0	0.02	0.049	0.825401
C	1	11	10.61	24.249	8.55e-07
SS:SD	1	0	0.31	0.713	0.398481
SS:DD	1	9	9.35	21.355	3.85e-06
SD:DD	1	56	55.63	127.094	< 2e-16
SS:E	1	0	0.01	0.019	0.889825
SD:E	1	47	47.46	108.442	< 2e-16
DD:E	1	65	64.57	147.536	< 2e-16
SS:C	1	0	0.02	0.056	0.813553
SD:C	1	15	15.22	34.784	3.76e-09
DD:C	1	15	14.74	33.671	6.65e-09
E:C	1	15	14.84	33.907	5.89e-09
SS:SD:DD	1	9	9.27	21.182	4.21e-06
SS:SD:E	1	0	0.01	0.019	0.889005
SS:DD:E	1	1	1.32	3.016	0.082482
SD:DD:E	1	0	0.04	0.081	0.776557
SS:SD:C	1	0	0.02	0.049	0.824964
SS:DD:C	1	6	5.89	13.460	0.000245
SD:DD:C	1	17	16.55	37.806	8.00e-10
SS:E:C	1	7	7.30	16.681	4.45e-05
SD:E:C	1	17	16.78	38.341	6.09e-10
DD:E:C	1	22	21.78	49.759	1.81e-12
SS:SD:DD:C	1	6	5.87	13.422	0.000250
SS:SD:E:C	1	7	7.47	17.056	3.65e-05
SD:DD:E:C	1	6	5.98	13.656	0.000220
Residuals	15719	6880	0.44		

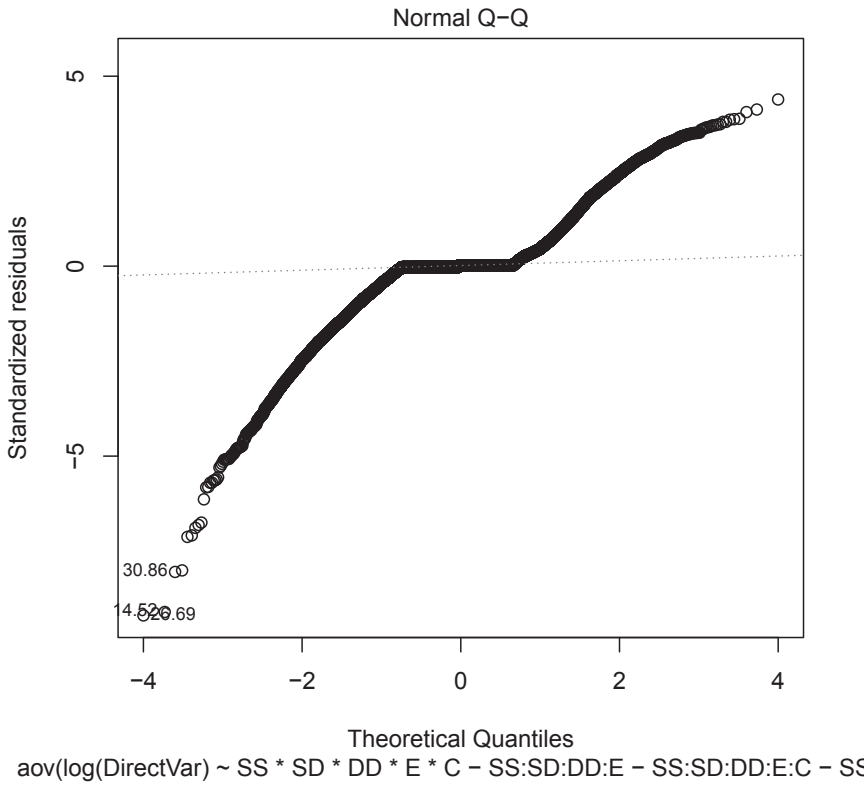

Figure 3.5: QQ Plot for Direct Error ANOVA After Selection With Removed Residuals

We can see similar results to before in which of the interactions are significant, and if we look at the QQ plot for this ANOVA shown in Figure 3.5, we can see that the same issues arise with long tails that seem to deviate from normality. A log transform for this model was also done, but the results look similar to the non-transformed version with deviating tails as well.

Instead of using the squared error as above, here we use the unmodified version of the estimated Direct effects variance. This allows us to investigate if the magnitude of the estimates is affected by the initial true values of the variances, rather than the error of the estimates. In Table 3.6 again we see that there are a lot of significant variables, probably due to the significance of several fourth order interaction terms. The QQ plot in Figure 3.6 that follows shows the same issues we continue to have, that of long tails deviating from normality, especially in the upper tailed case.

Table 3.6: Unmodified Direct Variance ANOVA After Selection

Effect	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SS	1	14.2	14.2	120.902	< 2e-16
SD	1	26.2	26.2	222.436	< 2e-16
DD	1	1582.9	1582.9	13464.199	< 2e-16
E	1	96.2	96.2	818.305	< 2e-16
C	1	8.3	8.3	70.813	< 2e-16
SS:SD	1	14.2	14.2	120.584	< 2e-16
SS:DD	1	3.8	3.8	32.451	1.24e-08
SD:DD	1	5.2	5.2	44.544	2.57e-11
SS:E	1	13.9	13.9	117.896	< 2e-16
SD:E	1	33.6	33.6	285.442	< 2e-16
DD:E	1	13.8	13.8	117.101	< 2e-16
SS:C	1	16.3	16.3	138.655	< 2e-16
SD:C	1	93.9	93.9	799.078	< 2e-16
DD:C	1	76.3	76.3	648.751	< 2e-16
E:C	1	65.5	65.5	556.747	< 2e-16
SS:SD:DD	1	3.8	3.8	32.609	1.15e-08
SS:SD:E	1	13.8	13.8	117.633	< 2e-16
SS:DD:E	1	16.5	16.5	140.545	< 2e-16
SD:DD:E	1	32.8	32.8	279.046	< 2e-16
SS:SD:C	1	16.3	16.3	138.757	< 2e-16
SS:DD:C	1	7.4	7.4	63.031	2.17e-15
SD:DD:C	1	0.5	0.5	3.871	0.0491
SS:E:C	1	7.2	7.2	61.233	5.39e-15
DD:E:C	1	12.9	12.9	109.914	< 2e-16
SS:SD:DD:E	1	16.6	16.6	140.779	< 2e-16
SS:SD:DD:C	1	7.4	7.4	63.055	2.14e-15
SS:SD:E:C	1	4.2	4.2	36.010	2.01e-09
SS:DD:E:C	1	10.2	10.2	86.756	< 2e-16
SD:DD:E:C	1	72.4	72.4	615.468	< 2e-16
Residuals	15960	1876.3	0.1		

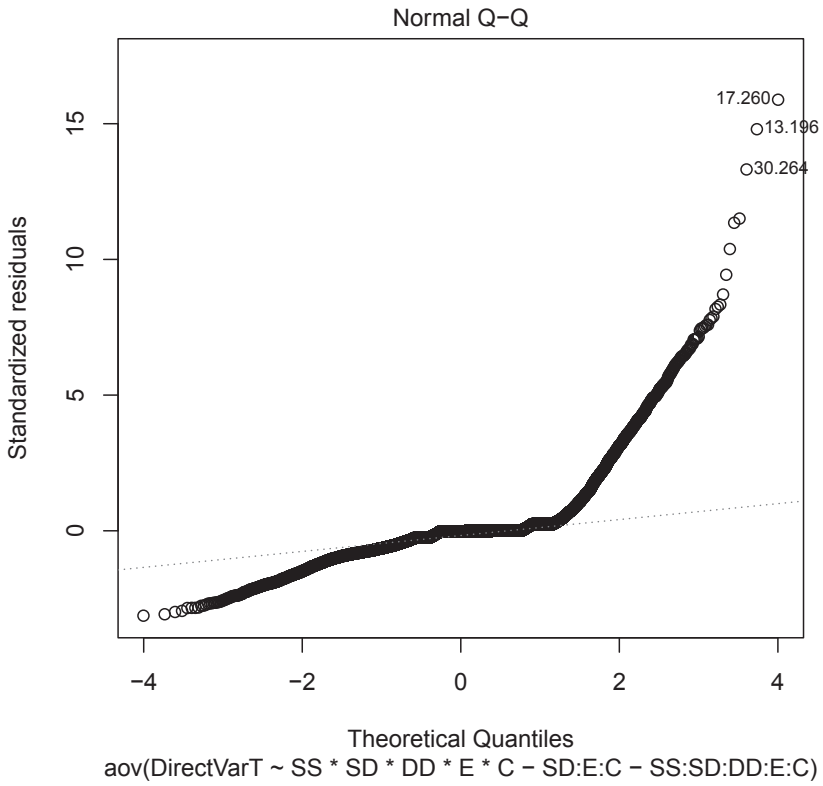

Figure 3.6: QQ Plot for Unmodified Direct Variance ANOVA After Selection

Since we have the same problem in the QQ plot as before, a log transform was performed as before. The results of doing this for the unmodified Direct variance are in Table 3.7 below. The fourth order interactions are still highly significant, and the QQ plot in Figure 3.7 shows the same deviation from normality, however in this case the lower tail seems to be more of an issue than the upper.

Table 3.7: Log Transform Unmodified Direct Variance ANOVA After Selection

Effect	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SS	1	103	103	163.29	< 2e-16
SD	1	0	0	0.14	0.709
DD	1	13210	13210	20871.82	< 2e-16
E	1	1435	1435	2267.42	< 2e-16
C	1	121	121	191.95	< 2e-16
SS:SD	1	103	103	162.90	< 2e-16
SS:DD	1	29	29	46.19	1.11e-11
SD:DD	1	70	70	111.11	< 2e-16
SS:E	1	64	64	101.16	< 2e-16
SD:E	1	429	429	677.90	< 2e-16
DD:E	1	415	415	655.05	< 2e-16
SS:C	1	89	89	141.22	< 2e-16
SD:C	1	649	649	1025.46	< 2e-16
DD:C	1	642	642	1014.42	< 2e-16
E:C	1	472	472	745.93	< 2e-16
SS:SD:DD	1	30	30	46.63	8.88e-12
SS:SD:E	1	64	64	100.78	< 2e-16
SS:DD:E	1	71	71	112.85	< 2e-16
SD:DD:E	1	944	944	1491.55	< 2e-16
SS:SD:C	1	89	89	141.22	< 2e-16
SS:DD:C	1	45	45	70.73	< 2e-16
SD:DD:C	1	50	50	78.72	< 2e-16
SS:E:C	1	57	57	90.42	< 2e-16
DD:E:C	1	72	72	113.33	< 2e-16
SS:SD:DD:E	1	72	72	113.32	< 2e-16
SS:SD:DD:C	1	45	45	70.70	< 2e-16
SS:SD:E:C	1	86	86	136.61	< 2e-16
SS:DD:E:C	1	61	61	96.96	< 2e-16
SD:DD:E:C	1	432	432	682.90	< 2e-16
Residuals	15960	10101	1		

3.3.2 Social Variance

The following is using the social, or indirect variance as the response variable rather than the direct variance as above. Figure 3.17 and the end of the chapter shows box plots for the estimates of the indirect variance for each of the 32 designs. The green line on the plot is a marker to show the true value for the variable's low value, while the red is representative of the high value. This plot isn't very helpful as there are a few large outliers, and since the estimated values are so small we can't really tell what's going on here. Figure 3.18 is the same plot with limited axes in order to cut out the large outliers zoom in on what's going on. We can see that in general the indirect variance is estimated more poorly than the direct one, with larger boxes and average estimates further away from the true values. We do see the same pattern as before in that the designs containing covariance do not have as many outliers, however even with this some designs preform poorly, namely designs 15 and 16, which are designs with high and low indirect variance and high direct and error variances. There are

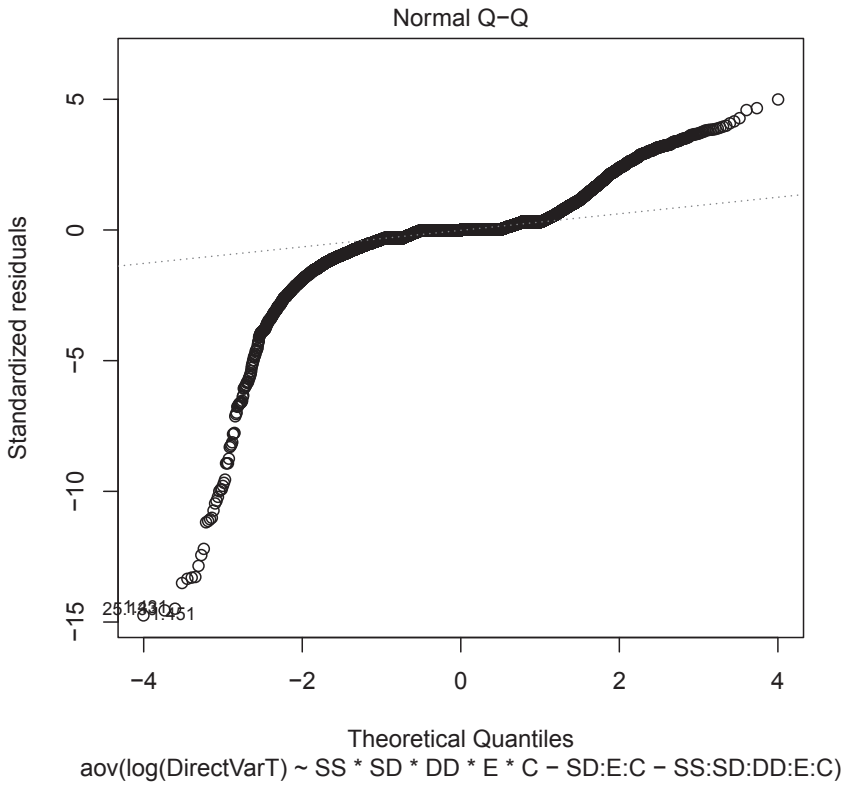

Figure 3.7: QQ Plot for Log Transform Unmodified Direct Variance ANOVA After Selection

also a few cases of a design estimating a value closer to the low value when it should be high, or vice versa, namely in cases 4, 7, 12, 20, 23, 27 and 32.

The ANOVA shown in Table 3.8 is for the model after selection, and it seems to show that there are many less significant factors than in the direct variance case. There is a mildly significant third order interaction, which forces many of the other variables to remain in the model. The most significant factor is the covariance, which is not unexpected. The QQ plot for the ANOVA in Figure 3.8 shows similar issues as before, however it is slightly better. Only the one tail is strongly deviating from normality, which may result in an acceptable log transform.

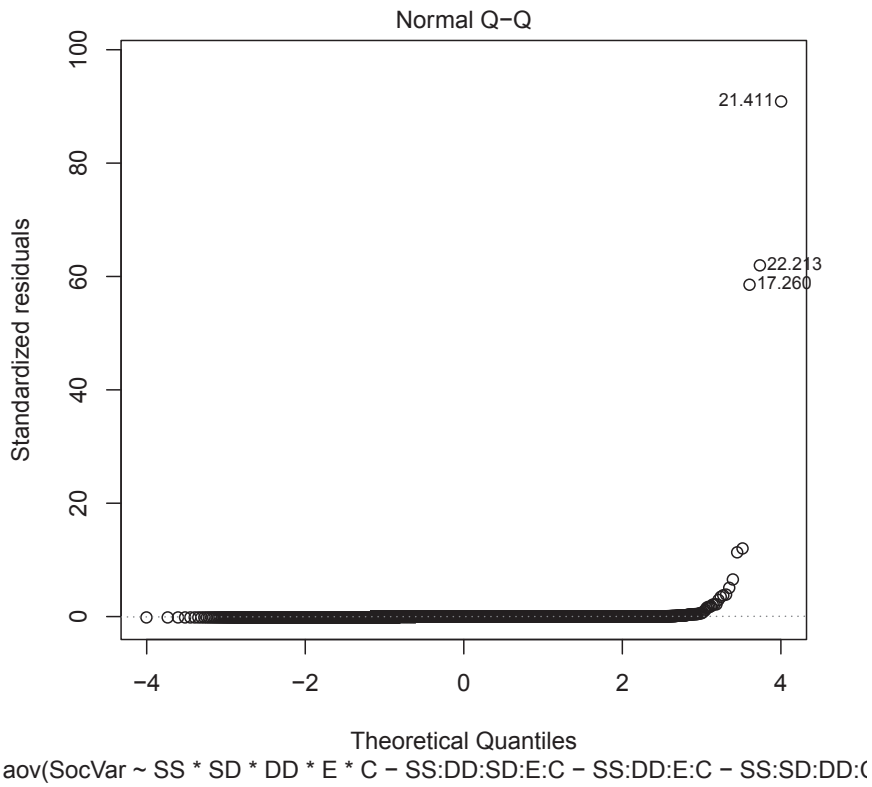

Figure 3.8: QQ Plot for Social Variance Error ANOVA After Selection

Table 3.8: Social Variance Error ANOVA After Selection

Effect	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SS	1	0	0.4263	0.961	0.3269
SD	1	2	2.4816	5.595	0.0180
DD	1	0	0.0923	0.208	0.6483
E	1	2	1.6015	3.611	0.0574
C	1	1	0.9370	2.113	0.1461
SD:E	1	2	1.7193	3.877	0.0490
SD:C	1	1	1.0243	2.309	0.1286
E:C	1	1	1.4581	3.288	0.0698
SD:E:C	1	1	1.3540	3.053	0.0806
Residuals	15980	7088	0.4435		

The results of the log transform performed on the social variance, and then model selected is below in Table 3.9. The results are similar to the non-transformed version, with a significant third order interaction. However in this case variables are much more significant. The QQ plot for this transform shown in Figure 3.9 is better than any shown for the direct variance, although the tails still do not look ideal.

Table 3.9: Social Variance Log Transform Error ANOVA After Selection

Effect	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SS	1	5831	5831	757.557	< 2e-16
SD	1	5404	5404	702.072	< 2e-16
DD	1	75	75	9.773	0.00177
E	1	3291	3291	427.589	< 2e-16
C	1	2958	2958	384.331	< 2e-16
SD:E	1	5584	5584	725.470	< 2e-16
SD:C	1	1854	1854	240.904	< 2e-16
E:C	1	913	913	118.656	< 2e-16
SD:E:C	1	1059	1059	137.627	< 2e-16
Residuals	15980	122994	8		

As we did for the direct variance, here we preform an ANOVA on the unmodified estimate of the social variance, the results of which are in Table 3.10. This gives more significant factors than the previous error estimate variance due to additional significant third order interaction terms, in particular terms involving the error and cage effects show up more prominently here, which is what we would expect in this unmodified case. The QQ plot in Figure 3.10 shows the same one deviating tail indicating possible transformation.

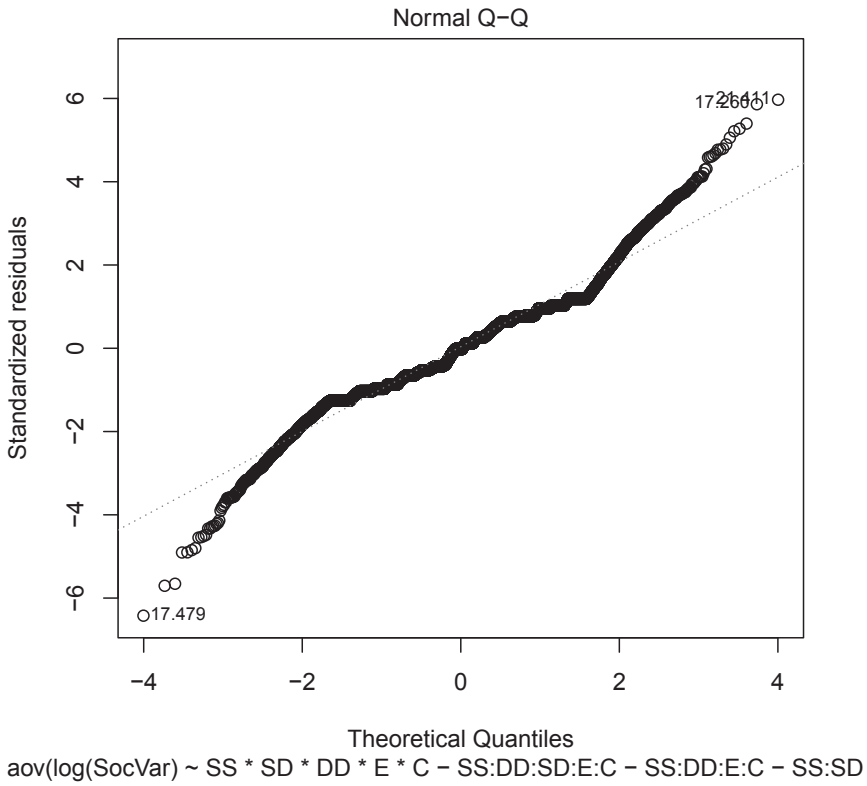

Figure 3.9: QQ Plot for Social Variance Log Transform Error ANOVA After Selection

Table 3.10: Unmodified Social ANOVA After Selection

Effect	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SS	1	0.02	0.0198	1.582	0.208445
SD	1	0.28	0.2841	22.717	1.89e-06
DD	1	0.06	0.0610	4.881	0.027172
E	1	0.01	0.0051	0.406	0.524122
C	1	0.22	0.2173	17.375	3.09e-05
SS:SD	1	0.02	0.0199	1.595	0.206567
SS:DD	1	0.00	0.0010	0.082	0.774102
SD:DD	1	0.44	0.4446	35.559	2.53e-09
SS:E	1	0.00	0.0048	0.383	0.536237
SD:E	1	0.69	0.6906	55.233	1.12e-13
DD:E	1	0.19	0.1900	15.197	9.72e-05
SS:C	1	0.01	0.0096	0.766	0.381427
SD:C	1	0.11	0.1148	9.183	0.002446
DD:C	1	0.15	0.1532	12.252	0.000466
E:C	1	0.30	0.2958	23.657	1.16e-06
SS:DD:C	1	0.04	0.0440	3.515	0.060819
SD:DD:C	1	0.32	0.3200	25.595	4.26e-07
DD:E:C	1	0.49	0.4905	39.229	3.87e-10
Residuals	15971	199.70	0.0125		

The log transform of the unmodified social variance estimates was done, and the ANOVA table after model selection is shown in Table 3.11. The same pattern appears in that now there are more significant terms, mostly due to a significant higher interaction. The QQ plot of this transform in Figure 3.11 shows fairly poor results. Both tails deviate strongly from normality rather than just the one as before.

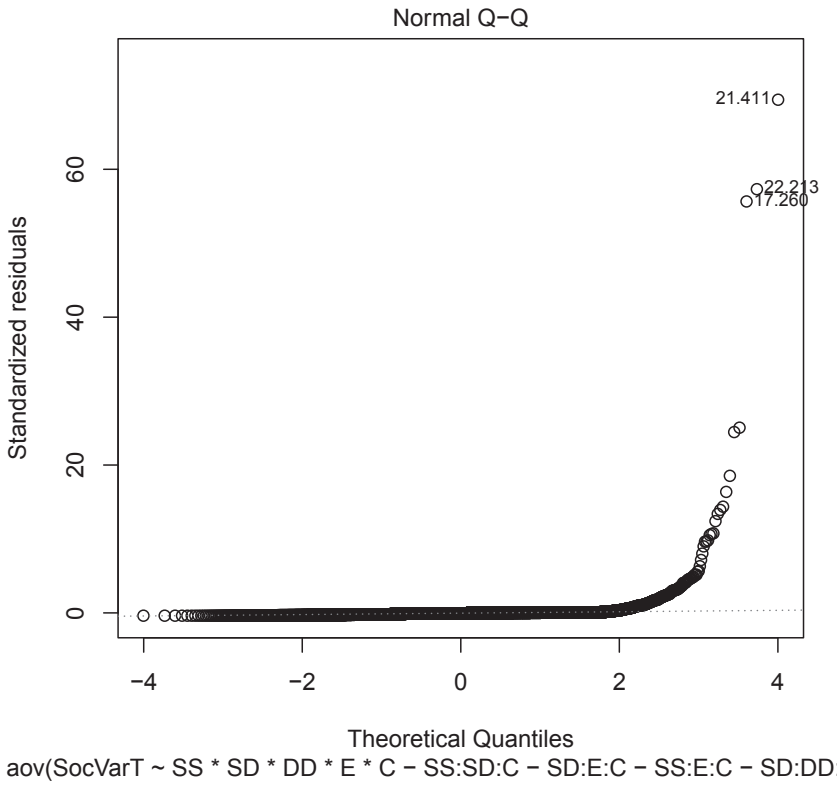

Figure 3.10: QQ Plot for Unmodified Social ANOVA After Selection

Table 3.11: Log Transform of Unmodified Social Variance ANOVA After Selection

Effect	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SS	1	3635	3635	1656.771	< 2e-16
SD	1	1300	1300	592.588	< 2e-16
DD	1	1822	1822	830.386	< 2e-16
E	1	1344	1344	612.741	< 2e-16
C	1	2809	2809	1280.138	< 2e-16
SS:SD	1	4561	4561	2078.747	< 2e-16
SS:DD	1	11	11	4.975	0.025734
SD:DD	1	2987	2987	1361.546	< 2e-16
SS:E	1	6	6	2.786	0.095098
SD:E	1	2561	2561	1167.185	< 2e-16
DD:E	1	764	764	348.228	< 2e-16
SS:C	1	40	40	18.007	2.21e-05
SD:C	1	1025	1025	467.388	< 2e-16
DD:C	1	27	27	12.222	0.000474
E:C	1	87	87	39.627	3.15e-10
SS:DD:C	1	30	30	13.623	0.000224
SD:DD:C	1	117	117	53.475	2.74e-13
SD:E:C	1	44	44	20.275	6.76e-06
DD:E:C	1	4223	4223	1924.858	< 2e-16
SD:DD:E:C	1	3763	3763	1715.187	< 2e-16
Residuals	15969	35037	2		

3.3.3 Covariance

Figure 3.19 at the end of the chapter shows the box plots containing the covariance estimates for the 32 designs. The green line on the plot is a marker to show the true value for the variable's low value, while the red is representative of the high value. There are some interesting things happening in this chart. Initially we see the same thing as in the previous two figures, that the designs that contained covariance in their initial parameters do not contain many outliers, with the estimates being much

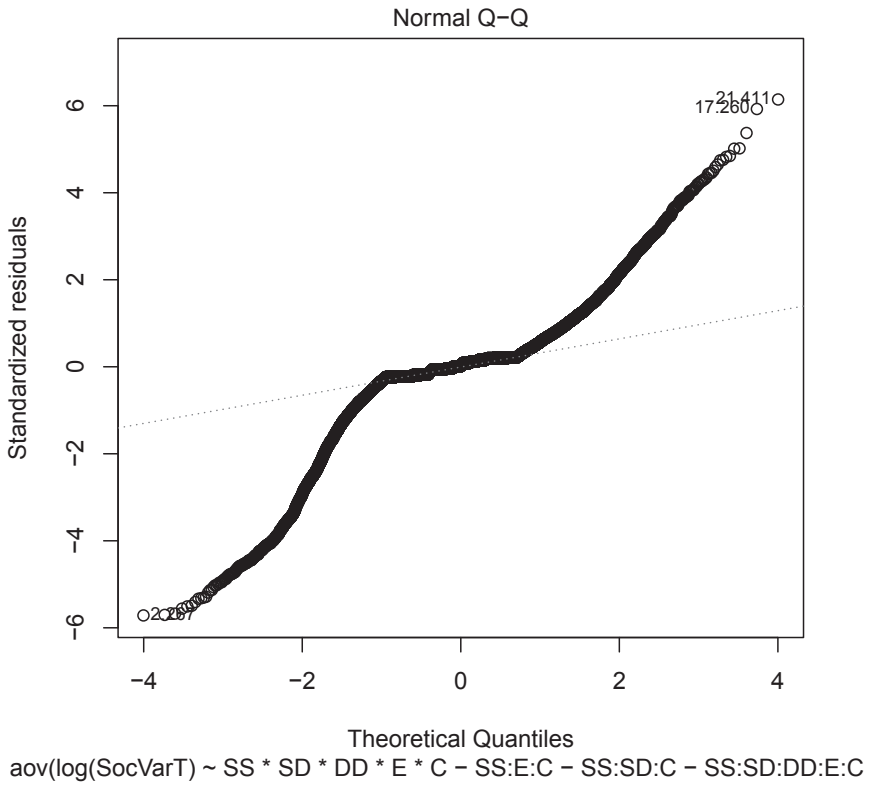

Figure 3.11: QQ Plot for Log Transform of Unmodified Social Variance ANOVA After Selection

more focused. However, interestingly whether the true generating covariance value was zero or not, the model seemed to predict values close to zero anyway, with only a few designs deviating much from zero.

The covariance between the two terms was modelled, starting initially with an error estimate of the squared value of the true value of the covariance subtracted from the estimated value of the covariance. The ANOVA results in Table 3.12 are much smaller than either of the social or direct variance models, with only one second order interaction term remaining significant after model selection. Interestingly the two most significant variables affecting the error of the estimate is the true value of the covariance itself, and the interaction term between the cage effect and the residual error. This is more in line with what we would expect in general, in that the accuracy of the estimate of the covariance depends largely on the actual value of the covariance, and the other sources of error. Unfortunately once again we obtain a poor QQ plot shown in Figure 3.12 with a large upper tail deviation.

Table 3.12: Covariance ANOVA Error After Selection

Effect	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SS	1	0.3	0.281	1.958	0.16176
SD	1	20.2	20.213	140.648	< 2e-16
DD	1	0.0	0.009	0.061	0.80477
E	1	0.5	0.520	3.621	0.05706
C	1	0.3	0.348	2.419	0.11990
E:C	1	1.5	1.515	10.541	0.00117
Residuals	15983	2297.0	0.144		

The log transform of the covariance error is performed and the results shown in Table 3.13. We obtain even less significant terms, with the error and cage interaction term no longer remaining significant, however what remains is all of the single terms, with each being highly significant. The QQ plot for this model in Figure 3.14 shows strong deviations in both tails.

Table 3.13: Log Transform of Covariance Error ANOVA After Selection

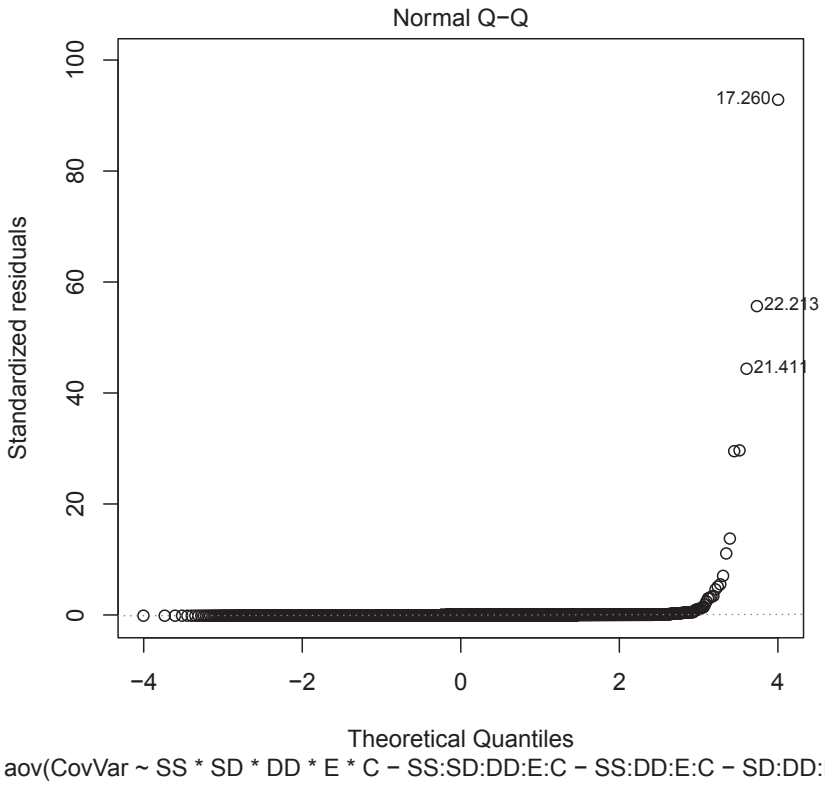

Figure 3.12: QQ Plot for Covariance ANOVA Error After Selection

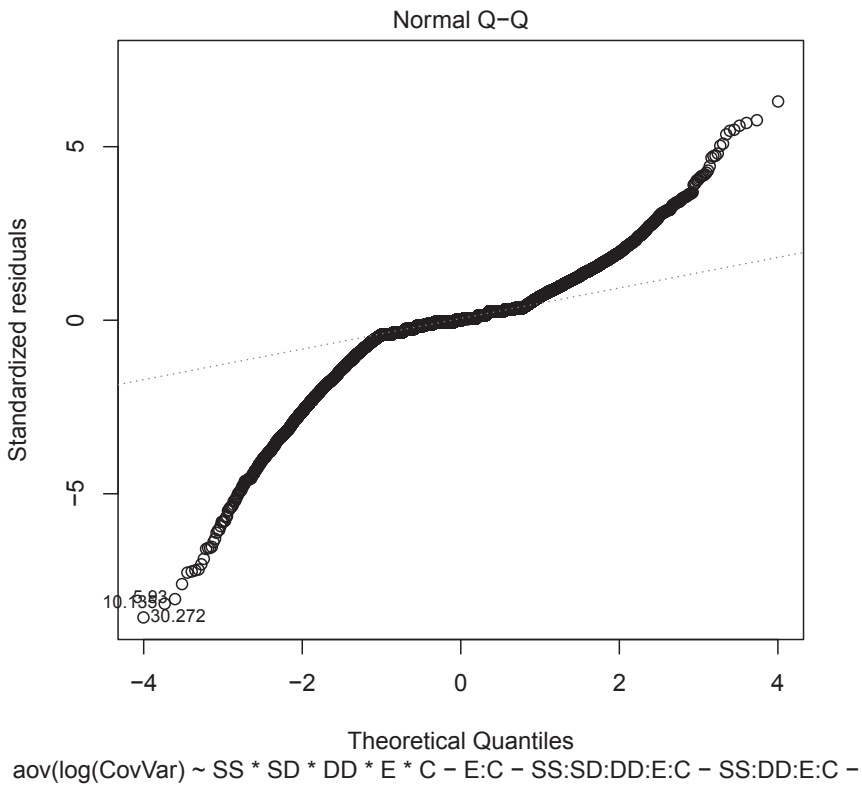

Figure 3.13: QQ Plot for Log Transform of Covariance Error ANOVA After Selection

Effect	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SS	1	1621	1621	364.81	< 2e-16
SD	1	185422	185422	41736.67	< 2e-16
DD	1	1289	1289	290.03	< 2e-16
E	1	53	53	11.88	0.000568
C	1	62	62	13.85	0.000198
Residuals	15984	71012	4		

Here we show the ANOVA performed on the unmodified covariance estimates after model selection was completed in Table 3.14. There are many more significant terms, especially due to the significance of several of the fourth order interactions. The significance of so many of the terms doesn't really tell us very much, especially since the QQ plot has a strongly deviating upper tail.

Table 3.14: Unmodified Covariance ANOVA After Selection

Effect	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SS	1	0.02	0.0228	2.597	0.10708
SD	1	0.04	0.0447	5.087	0.02412
DD	1	0.00	0.0012	0.131	0.71698
E	1	0.01	0.0087	0.993	0.31893
C	1	0.01	0.0054	0.616	0.43263
SS:SD	1	0.02	0.0229	2.606	0.10647
SS:DD	1	0.00	0.0043	0.493	0.48274
SD:DD	1	0.02	0.0196	2.233	0.13508
SS:E	1	0.00	0.0005	0.059	0.80758
SD:E	1	0.08	0.0807	9.186	0.00244
DD:E	1	0.14	0.1378	15.679	7.54e-05
SS:C	1	0.00	0.0001	0.015	0.90120
SD:C	1	0.03	0.0323	3.677	0.05519
DD:C	1	0.02	0.0234	2.668	0.10242
E:C	1	0.85	0.8462	96.290	$<$ 2e-16
SS:SD:DD	1	0.00	0.0043	0.494	0.48233
SS:SD:E	1	0.00	0.0005	0.061	0.80429
SS:DD:E	1	0.04	0.0351	3.992	0.04572
SD:DD:E	1	0.05	0.0539	6.138	0.01324
SD:DD:C	1	0.15	0.1480	16.839	4.09e-05
SD:E:C	1	1.01	1.0147	115.465	$<$ 2e-16
DD:E:C	1	0.53	0.5299	60.304	8.62e-15
SS:SD:DD:E	1	0.03	0.0350	3.982	0.04600
SD:DD:E:C	1	0.76	0.7620	86.713	$<$ 2e-16
Residuals	15965	140.30	0.0088		

We include the log transform of the unmodified covariance ANOVA results as well in Table 3.15. There are slightly fewer significant variables, with higher levels of significance, but the QQ plot in Figure 3.15 has two strongly deviating tails.

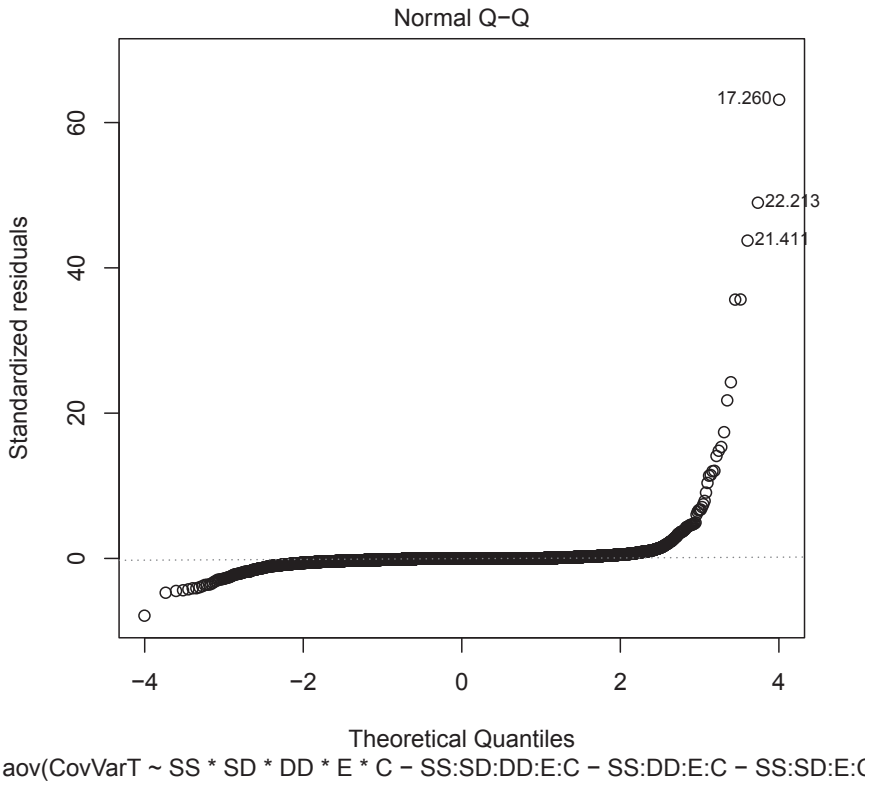

Figure 3.14: QQ Plot for Unmodified Covariance ANOVA After Selection

Table 3.15: Log Transform of Unmodified Covariance ANOVA After Selection

Effect	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SS	1	189	188.8	169.477	< 2e-16
SD	1	90	90.1	80.862	< 2e-16
DD	1	2719	2719.1	2440.661	< 2e-16
E	1	366	366.1	328.622	< 2e-16
C	1	595	594.8	533.868	< 2e-16
SS:SD	1	200	200.3	179.810	< 2e-16
SS:DD	1	8	7.9	7.083	0.007796
SD:DD	1	527	527.0	473.014	< 2e-16
SD:E	1	339	338.8	304.075	< 2e-16
DD:E	1	446	446.3	400.560	< 2e-16
SS:C	1	7	6.6	5.911	0.015073
SD:C	1	718	718.0	644.514	< 2e-16
DD:C	1	29	29.4	26.392	2.85e-07
E:C	1	14	14.5	12.973	0.000318
SS:SD:DD	1	15	15.2	13.603	0.000227
SS:DD:E	1	10	10.0	8.957	0.002772
SS:E:C	1	13	13.1	11.722	0.000621
Residuals	8011	8925	1.1		

3.3.4 Additional Analysis

There were other forms of analysis done, and for brevity we have excluded the graphs, but a brief description of some of the techniques used is useful. Ideally what we care about in this problem is which of the variance components are important in determining the estimate error, a CART method was performed on the same data, as CART methods are good at identifying important variables due to the splitting nature of trees. This method however tended to give non-informative trees as a result, with the most important result being one split related to the covariance. This is an expected result in that when the covariance value is non-zero, the estimated error tends to be larger. Other used methods that did not yield useful results include bump hunting with PRIM, and generalised linear mixed modelling. In particular

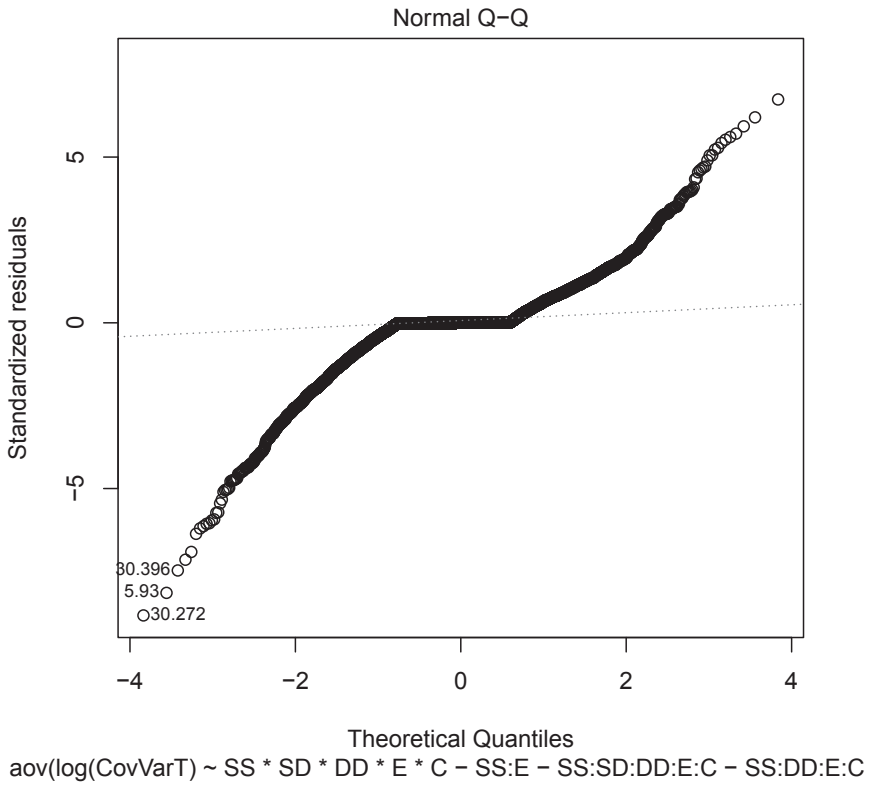

Figure 3.15: QQ Plot for Log Transform of Unmodified Covariance ANOVA After Selection

GLM's were preformed using long tailed error distributions in order to more closely match the long tailed distributions we obtained in the ANOVA results. This was done both using model selection techniques and by just directly taking the variables deemed significant by the appropriate ANOVA and modeling it using a GLM instead. A table of the p-values in order to compare the three approaches can be found in the appendix as Table 1, Table 2 and Table 3, however they seem to show roughly similar results in which variables are significant, and their significance values.

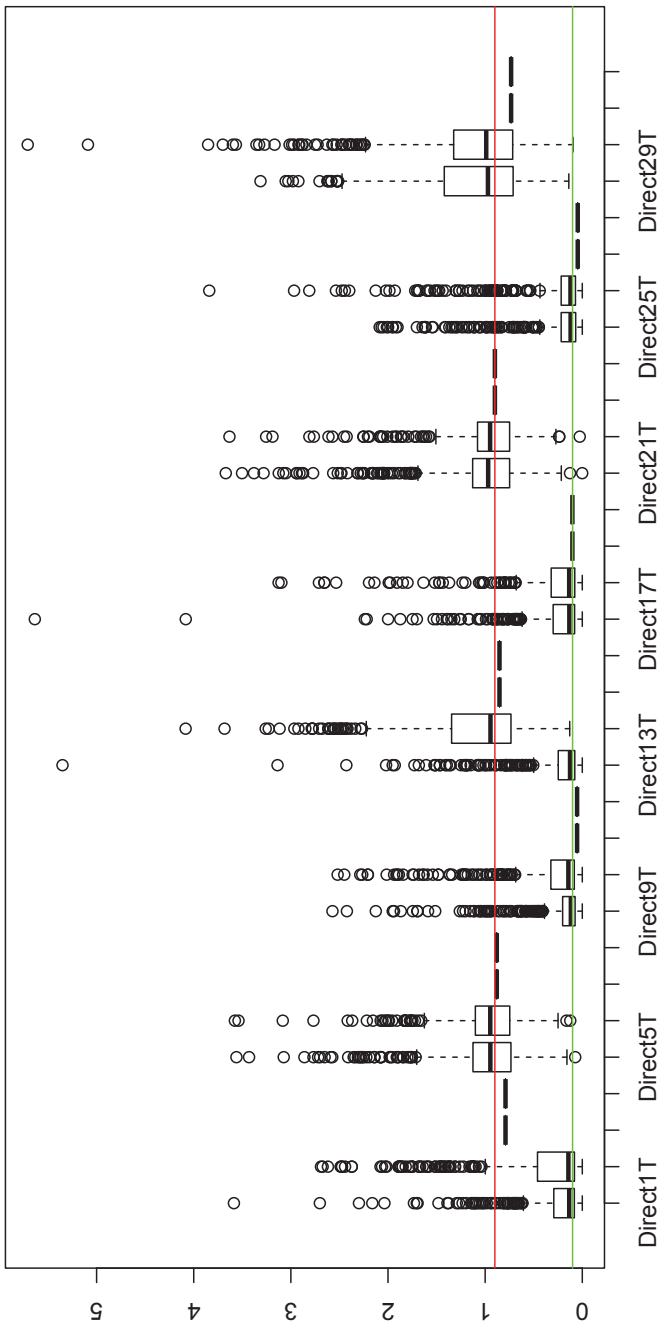

Figure 3.16: Box Plots of Direct Variance Estimates

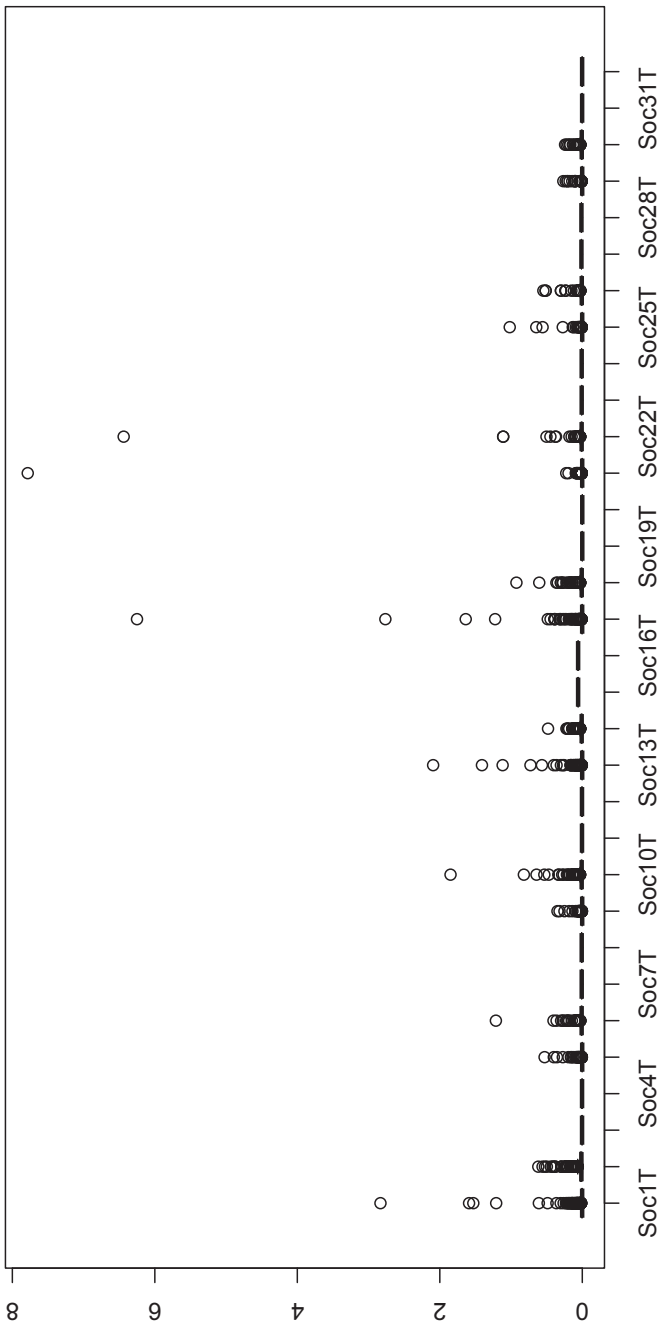

Figure 3.17: Box Plots of Social Variance Estimates

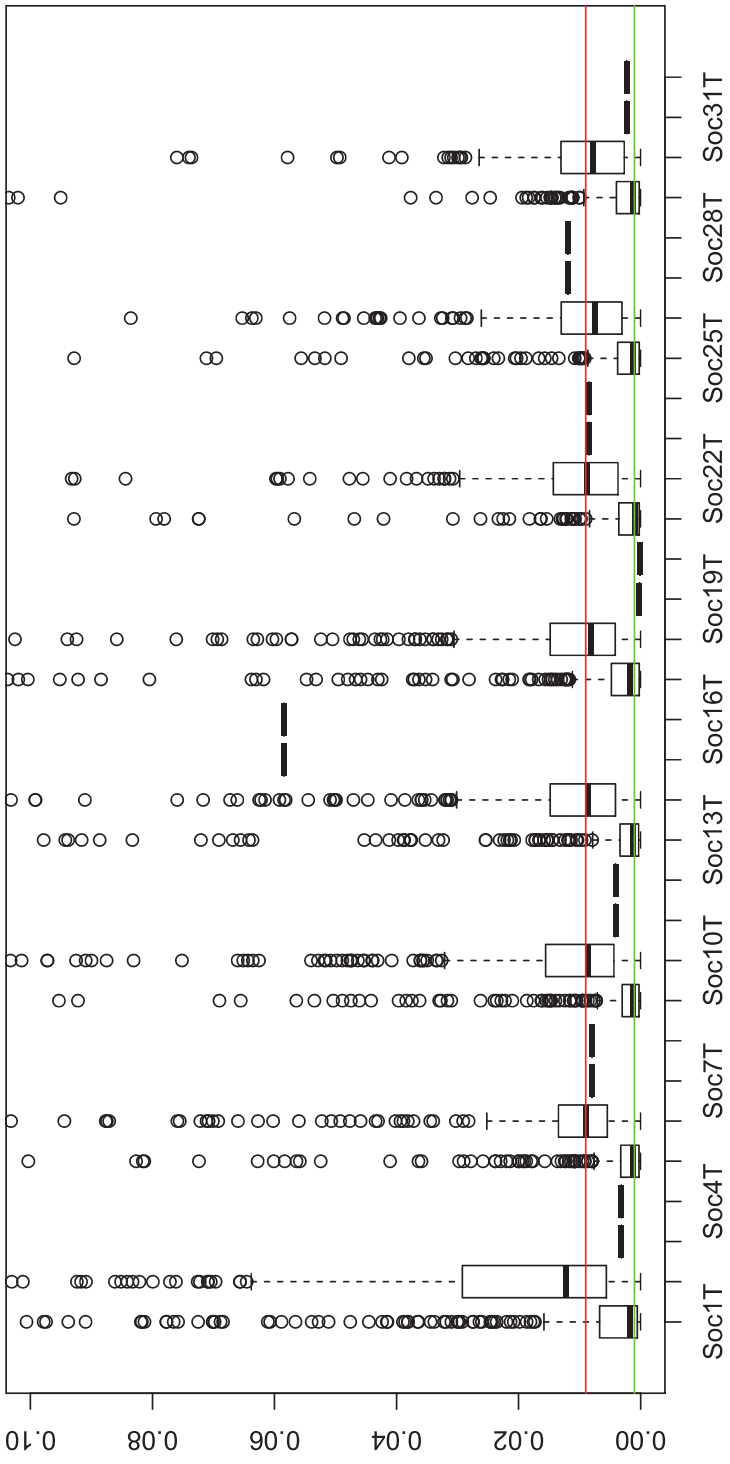

Figure 3.18: Zoomed Box Plots of Social Variance Estimates

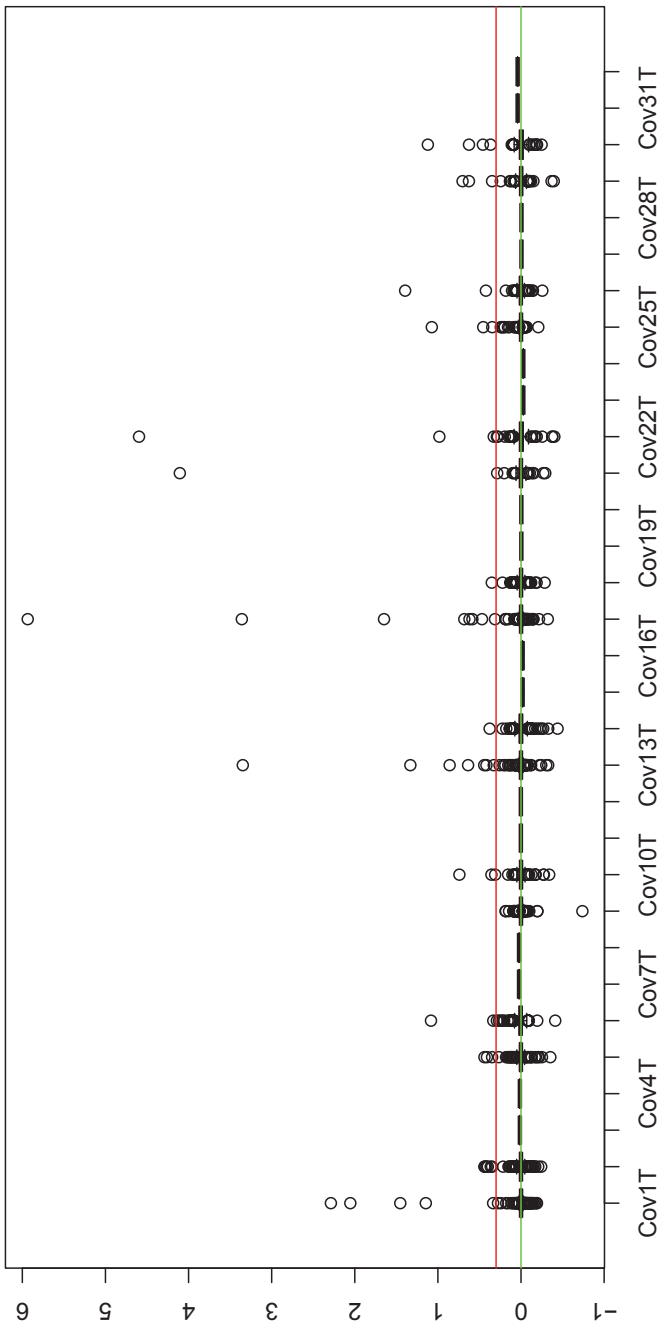

Figure 3.19: Box Plots of Covariance Estimates

Chapter 4

Conclusions

4.1 Summary

4.1.1 Boxplots

The box plots indicate that the direct genetic variance is often estimated more accurately using this model than the indirect social one is, however there are a few situations for both of these parameters where the estimation seems to be inaccurate. This is not unexpected in a biological context, as the effects of direct genetics are understood, and the social or indirect effects are more complicated in that there may be many things that can fall under the umbrella of social effects, rather than the straightforward nature of direct effects. This is also shown in that there are more designs in which the social variance was estimated inaccurately than there are designs where the direct was. In any case, the use of this model in the future should be monitored for those situations in which the estimations seem to be off. The covariance parameter is often estimated as zero or close to zero even if the true value is non-zero, and we do have the cases where the variability of the estimates seem to collapse when we have a non-zero covariance as a parameter. This may be due to the nature of the data generation in that the value chosen as non-zero value may have been too high. This could result in a convergent solution in the REML process in estimation. This means that any conclusions about the cases where the covariance parameter was set to a non-zero value should be taken with a grain of salt, and may not be informative at all. However the parameter sets with no covariance present still offer valuable insight.

4.1.2 Anova Results

In most of the ANOVA results there are a large number of significant interaction terms, showing that the nature of the proposed social model is complex. That is,

the accuracy of this models estimation is affected by a large variation of possible initial states, or different sets of true values for the various parameters. Something important to note however is that a large number of terms were significant, with very small p-values, this is very likely influenced strongly by the large amount of data points, giving a very large residual degrees of freedom, leading to very small MSE values. The co-variance ANOVA, and in particular the log transformed version had no significant interaction terms. Unsurprisingly, the most significant factor was the covariance itself, and since the two values used in the simulation were 0 and a non-zero value we see that the most significant factor in the accuracy of the estimation of the covariance is weather or not a covariance exists between the social and direct variances. However it is of more importance than the true social variance when estimating the social variance, while the true direct variance has more significance than the covariance when estimating the direct variance. Another term that seems to appear with large significance for all three of the variance terms is the interaction between the fixed cage effect, and the error term. The error term has a higher significance in the social variance and covariance ANOVA's, while the cage effect has a higher significance in the direct variance ANOVA, which is also an interesting result. In general, the most complicated term to estimate seems to be the direct variance, with a large number of factors affecting its accuracy, and patterns that appear in the terms which are significant for the covariance and social variance ANOVA's do not appear to exist in the same way in the direct variance ANOVA's. We did see in the boxplots that the direct variance estimate tends to be more accurate than the social variance or covariance estimates, but according to the ANOVA results the direct variance estimate has more factors influencing it than the social variance.

4.1.3 QQ Plots

While we can see that the QQ plots provided in the previous section are not what one would consider ideal, we need to consider the nature of the analysis we are doing. We are first using BLUP to estimate the variances contained in the social model, that is the direct effect variance, social effect variance, and covariance between these two. We repeat this several times to obtain a data set, then use ANOVA to analyse this data. Thus we are estimating the variance of a set of variances. This would

indicate that the long tailed distributions we seem to obtain in all of the analyses is not an unexpected result. This, combined with the large number of data points would indicate that despite the long tailed distributions we see, we would expect that the ANOVA results would be trustworthy.

4.2 Future Work

This work was done with the intention of examining the proposed social model for the purposes of its use with a particular experiment performed at Dalhousie University, and so its scope was limited. While its specific use in terms of aquaculture is generally important, the nature of the design could easily be changed for future research into this topic. In particular examining how the model behaves with data generated from a balanced complete block design to see how it's accuracy is affected under ideal conditions. Many possibilities exist for this including random allocation of families if an incomplete design were to be used. There were some other simulated designs, as well as a small analysis of real data during the course of the researched performed for this work, however it was mostly used to tune the generation process, and to determine which of the available BLUP tools would be most appropriate for examining this problem. In general, more exploration of this model in an aquaculture setting could be useful, where the generated data does not attempt to match a particular experimental design, or the use of a different BLUP analysis tool to compare and contrast the results to see if similar conclusions are made. The other direction this could be taken in is a more specific one, in that the problems with certain situations for estimating the social or direct variances could be explored more, to see if this is repeatable, or more situations where variance estimation is poor could be found. The other interesting question is about the nature of covariance complication mentioned previously. Rerunning simulations with different values for the covariance parameter could offer insight into the possible issue that covariance seems to be underestimated in most cases, and to see if the strange situations where estimate variability seems to collapse continues to occur when smaller non-zero values of a generating covariance were to be used.

Appendices

Table 1: Direct Variance Pval Comparison

Term	ANOVA	GLM Selected	GLM Copy
SS	0.7057	3.16E-06	1.52E-06
SD	< 2e-16	< 2e-16	< 2e-16
DD	0.8359	0.2619	0.127785
E	0.5603	0.2165	0.000359
C	1.99E-07	0.09521	0.021905
SS:SD	-	0.05094	
SS:DD	5.69E-05	1.06E-07	1.20E-07
SD:DD	< 2e-16	< 2e-16	0.357479
SS:E	-	0.27193	-
SD:E	< 2e-16	< 2e-16	2.00E-16
DD:E	< 2e-16	0.00115	1.90E-08
SS:C	-	5.49E-05	-
SD:C	7.59E-08	< 2e-16	2.00E-16
DD:C	1.32E-09	0.8456	0.031467
E:C	1.35E-09	0.68552	0.010965
SS:SD:DD	0.0261	-	0.001461
SS:SD:E	-	-	-
SS:DD:E	-	0.07594	-
SD:DD:E	-	< 2e-16	-
SS:SD:C	-	0.6455	-
SS:DD:C	0.0384	0.00229	0.046196
SD:DD:C	1.21E-08	< 2e-16	< 2e-16
SS:E:C	-	0.03829	-
SD:E:C	9.21E-09	< 2e-16	2.00E-16
DD:E:C	1.64E-13	0.02033	0.003304
SS:SD:DD:E	-	-	-
SS:SD:DD:C	0.0386	-	0.038419
SS:SD:E:C	-	-	-
SS:DD:E:C	-	-	-
SD:DD:E:C	0.0206	< 2e-16	2.00E-16
SS:SD:DD:E:C	-	-	-

Table 2: Social Variance Pval Comparison

Term	ANOVA	GLM Selected	GLM Copy
SS	0.3269	0.00434	0.13465
SD	0.018	0.27041	0.00457
DD	0.6483	0.0008	0.49259
E	0.0574	0.01355	0.116
C	0.1461	0.01657	0.00036
SS:SD	-	0.28455	-
SS:DD	-	0.00372	-
SD:DD	-	0.22415	-
SS:E	-	0.00801	-
SD:E	0.049	0.12234	0.00202
DD:E	-	0.00036	-
SS:C	-	0.12919	-
SD:C	0.1286	0.02915	0.20446
DD:C	-	0.00042	-
E:C	0.0698	0.02216	0.00367
SS:SD:DD	-	0.85276	-
SS:SD:E	-	0.14864	-
SS:DD:E	-	0.00102	-
SD:DD:E	-	0.03873	-
SS:SD:C	-	0.01272	-
SS:DD:C	-	0.07293	-
SD:DD:C	-	0.008136	-
SS:E:C	-	0.022063	-
SD:E:C	0.0806	0.01745	0.03554
DD:E:C	-	0.00067	-
SS:SD:DD:E	-	0.60187	-
SS:SD:DD:C	-	0.01234	-
SS:SD:E:C	-	0.006	-
SS:DD:E:C	-	0.07784	-
SD:DD:E:C	-	0.00118	-
SS:SD:DD:E:C	-	0.00427	-

Table 3: Covariance Pval Comparison

Term	ANOVA	GLM Selected	GLM Copy
SS	0.16176	0.00134	0.68994
SD	< 2e-16	0.00095	0.00051
DD	0.80477	4.99E-05	0.94358
E	0.05706	0.00247	0.60621
C	0.1199	0.18233	0.34577
SS:SD	-	0.00135	-
SS:DD	-	0.00048	-
SD:DD	-	5.69E-05	-
SS:E	-	0.00699	-
SD:E	-	0.00247	-
DD:E	-	3.83E-05	-
SS:C	-	-	-
SD:C	-	0.18036	-
DD:C	-	0.0013	-
E:C	0.00117	0.15238	0.38587
SS:SD:DD	-	0.00049	-
SS:SD:E	-	0.00704	-
SS:DD:E	-	0.00118	-
SD:DD:E	-	4.31E-05	-
SS:SD:C	-	-	-
SS:DD:C	-	-	-
SD:DD:C	-	0.00154	-
SS:E:C	-	-	-
SD:E:C	-	0.15312	-
DD:E:C	-	0.00268	-
SS:SD:DD:E	-	0.0012	-
SS:SD:DD:C	-	-	-
SS:SD:E:C	-	-	-
SS:DD:E:C	-	-	-
SD:DD:E:C	-	0.00314	-
SS:SD:DD:E:C	-	-	-

Bibliography

[1] Rob Bergsma, Egbert Kanis, Egbert Frank Knol, and Piter Bijma. The contribution of social effects to heritable variation in finishing traits of domestic pigs (sus scrofa). *Genetics*, 178(3):1559–1570, 2008.

[2] Piter Bijma. Estimating indirect genetic effects: precision of estimates and optimum designs. *Genetics*, 186(3):1013–1028, 2010.

[3] Piter Bijma, William M Muir, and Johan AM Van Arendonk. Multilevel selection 1: quantitative genetics of inheritance and response to selection. *Genetics*, 175(1):277–288, 2007.

[4] Tim Clutton-Brock. Breeding together: kin selection and mutualism in cooperative vertebrates. *Science*, 296(5565):69–72, 2002.

[5] JV Craig and WM Muir. Group selection for adaptation to multiple-hen cages: beak-related mortality, feathering, and body weight responses. *Poultry Science*, 75(3):294–302, 1996.

[6] Noel Cressie and Soumendra Nath Lahiri. The asymptotic distribution of reml estimators. *Journal of multivariate analysis*, 45(2):217–233, 1993.

[7] Charles Darwin. On the origin of species by means of natural selection. 1859. *J. Murray, London*, 1991.

[8] L Otis Emik and Clair E Terrill. Systematic procedures for calculating inbreeding coefficients. *Journal of Heredity*, 40(2):51–55, 1949.

[9] Food and Agriculture Organization of the United Nations. *The State of World Fisheries and Aquaculture, 2014*. Food & Agriculture Org, Rome, 2014.

[10] Steven A Frank. *Foundations of social evolution*. Princeton University Press, 1998.

[11] Arthur S Goldberger. Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association*, 57(298):369–375, 1962.

[12] Bruce Griffing. Selection in reference to biological groups i. individual and group selection applied to populations of unordered groups. *Australian Journal of Biological Sciences*, 20(1):127–140, 1967.

[13] William D Hamilton. The genetical evolution of social behaviour. ii. *Journal of theoretical biology*, 7(1):17–52, 1964.

[14] DA Harville and TP Callanan. Computational aspects of likelihood-based inference for variance components. In *Advances in statistical methods for genetic improvement of livestock*, pages 136–176. Springer, 1990.

[15] Charles R Henderson. Estimation of genetic parameters (abstract). *The Annals of Mathematical Statistics*, 21:309–310, 1950.

[16] Charles R Henderson. Selection index and expected genetic advance. *Statistical genetics and plant breeding*, 982:141–163, 1963.

[17] Charles R Henderson. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, pages 423–447, 1975.

[18] Charles R Henderson. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics*, pages 69–83, 1976.

[19] CR Henderson and RL Quaas. Multiple trait evaluation using relatives' records. *Journal of Animal Science*, 43(6):1188–1197, 1976.

[20] Just Jensen and IL Mao. Transformation algorithms in analysis of single trait and of multitrait models with equal design matrices and one random factor per trait: a review. *Journal of Animal Science*, 66(11):2750–2761, 1988.

[21] M Jobling. Growth studies with fish - overcoming the problems of size variation. *Journal of Fish Biology*, 22(2):153–157, 1983.

[22] Ilan Karplus. Social control of growth in macrobrachium rosenbergii (de man): a review and prospects for future research. *Aquaculture Research*, 36(3):238–254, 2005.

[23] Laurent Keller. *Levels of selection in evolution*. Princeton University Press, 1999.

[24] Michael Lynch, Bruce Walsh, et al. *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, MA, 1998.

[25] Joel W McGlothlin and Edmund D Brodie III. How to measure indirect genetic effects: The congruence of trait-based and variance-partitioning approaches. *Evolution*, 63(7):1785–1795, 2009.

[26] K Meyer. Maximum likelihood estimation of variance components for a multivariate mixed model with equal design matrices. *Biometrics*, pages 153–165, 1985.

[27] K Meyer. Approximate accuracy of genetic evaluation under an animal model. *Livestock Production Science*, 21(2):87–100, 1989.

[28] K Meyer. Estimation of genetic parameters. In W.G Hill and T.F.C Mackay, editors, *Evolution and Animal Breeding*, pages 161–167. CAB International, 1989.

[29] K Meyer. The reml wombat homepage. `http://didgeridoo.une.edu.au/km/wombat.php`, 2014. [Online; accessed 26-November-2014].

[30] Timothy A Mousseau and Charles W Fox. *Maternal effects as adaptations*. Oxford University Press, 1998.

[31] Raphael A Mrode. *Linear models for the prediction of animal breeding values*. Cabi, 2014.

[32] S Msangi, M Kobayashi, M Batka, S Vannuccini, MM Dey, and JL Anderson. Fish to 2030: Prospects for fisheries and aquaculture. *World Bank Report*, (83177-GLB), 2013.

[33] William M Muir. Incorporation of competitive effects in forest tree or animal breeding programs. *Genetics*, 170(3):1247–1259, 2005.

[34] RL Quaas. Computing the diagonal elements and inverse of a large numerator relationship matrix. *Biometrics*, pages 949–953, 1976.

[35] RL Quaas and EJ Pollak. Modified equations for sire models with groups. *Journal of Dairy Science*, 64(9):1868–1872, 1981.

[36] LR Schaeffer. Estimation of variances and covariances within the allowable parameter space. *Journal of Dairy Science*, 69(1):187–194, 1986.

[37] Shayle R Searle, George Casella, and Charles E McCulloch. *Variance components*, volume 391. John Wiley & Sons, 2009.

[38] TJ Sweeting. Uniform asymptotic normality of the maximum likelihood estimator. *The Annals of Statistics*, pages 1375–1381, 1980.

[39] JF Taylor, B Bean, CE Marshall, and JJ Sullivan. Genetic and environmental components of semen production traits of artificial insemination holstein bulls. *Journal of Dairy Science*, 68(10):2703–2722, 1985.

[40] R Thompson, WG Hill, et al. Univariate reml analyses for multivariate data with the animal model. In *Proceedings of the 4th World Congress on Genetics applied to Livestock Production, Edinburgh 23-27 July 1990. XIII. Plenary lectures, molecular genetics and mapping, selection, prediction and estimation.*, pages 484–487, 1990.

[41] B Tier. Computing inbreeding coefficients quickly. *Genetics Selection Evolution*, 22(4):419–430, 1990.

[42] RL Willham. The covariance between relatives for characters composed of components contributed by related individuals. *Biometrics*, pages 18–27, 1963.

[43] O Wilson Edward. Sociobiology: the new synthesis. *Cambridge, MA: Belknap*, 1975.

[44] Jason B Wolf, Edmund D Brodie III, James M Cheverud, Allen J Moore, and Michael J Wade. Evolutionary consequences of indirect genetic effects. *Trends in Ecology & Evolution*, 13(2):64–69, 1998.

[45] Sewall Wright. Coefficients of inbreeding and relationship. *American Naturalist*, pages 330–338, 1922.