**University of São Paulo
“Luiz de Queiroz” College of Agriculture**

**Modeling strategies for complex hierarchical and overdispersed data in the
life sciences**

Izabela Regina Cardoso de Oliveira

Thesis submitted in fulfillment of the requirements
for the joint degree of Doctor in Science. Areas of
concentration: Statistics and Agricultural
Experimentation at ESALQ, and Statistics at U Hasselt

**Piracicaba
2014**

De transnationale Universiteit Limburg is een uniek samenwerkingsverband van twee universiteiten uit twee landen: de Universiteit Hasselt en Maastricht University.

De opleidingen informatica/kennistechnologie, statistiek en biomedische wetenschappen/moleculaire levenswetenschappen zijn reeds ondergebracht in dit samenwerkingsverband. De bacheloropleiding in de rechten is een gezamenlijk initiatief van de Universiteit Hasselt, Maastricht University en de KULeuven. Ook in andere wetenschapsdomeinen wordt gezocht naar samenwerking met andere universiteiten.

www.uhasselt.be
Universiteit Hasselt
Martelarenlaan 42 | BE-3500 Hasselt
T + 32(0)11 26 81 11

www.maastrichtuniversity.nl
P.O. Box 616 | NL- 6200 MD Maastricht
T + 31(0)43 388 2222

Modeling strategies for complex hierarchical and overdispersed data in the life sciences

Izabela Regina Cardoso de Oliveira

Maastricht University

universiteit
hasselt
KNOWLEDGE IN ACTION

2014 | Faculty of Sciences

DOCTORAL DISSERTATION

Modeling strategies for complex hierarchical and overdispersed data in the life sciences

Doctoral dissertation submitted to obtain the degree of doctor of science: statistics, to be defended by

Izabela Regina Cardoso de Oliveira

Promoters: Prof. Dr. Geert Molenberghs
Dr. Carlos Tadeu dos Santos Dias

Izabela Regina Cardoso de Oliveira
Bachelor in Business Administration

**Modeling strategies for complex hierarchical and overdispersed data in the
life sciences**

Advisers:
Prof. Dr. **CARLOS TADEU DOS SANTOS DIAS**
Prof. Dr. **GEERT MOLENBERGHS**

Thesis submitted in fulfillment of the requirements for
the joint degree of Doctor in Science. Areas of
concentration: Statistics and Agricultural
Experimentation at ESALQ, and Statistics at U Hasselt

**Piracicaba
2014**

**Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP**

Oliveira, Izabela Regina Cardoso de
Modeling strategies for complex hierarchical and overdispersed data in the
life sciences / Izabela Regina Cardoso de Oliveira.- - Piracicaba, 2014.
79 p: il.

Tese (Doutorado) - - Escola Superior de Agricultura "Luiz de Queiroz", 2014.

1. Modelo combinado 2. Distribuição gama 3. Modelo linear misto generalizado
4. Herdabilidade 5. Variâncias negativas 6. Distribuição Poisson 7. Efeito aleatório
8. Distribuição Weibull I. Título

CDD 519.5
O48m

"Permitida a cópia total ou parcial deste documento, desde que citada a fonte -O autor"

To the memory of my beloved grandmother

ACKNOWLEDGMENTS

I want to take this opportunity to express my gratitude to all people and institutions that make the realization of this work possible. This work was mainly financed by CNPq, National Council for Scientific and Technological Development, through grants in Brazil and abroad. I gratefully acknowledge this support. I also thank the CPG of ESALQ and Research Department of U Hasselt for all efforts to establish the agreement between both universities.

I am very grateful to my supervisors in this research. Thank you, Carlos Tadeu dos Santos Dias, for guidance and for respecting my time and my choices. Clarice Garcia Borges Demétrio, thanks for the ongoing encouragement, important advices and for the care. Geert Molenberghs, thank you for the warm welcome in Belgium, for the guidance, for always motivating me with your enthusiasm and for the valuable discussions about statistics, history, culture, climate and languages.

I especially thank my family, Tadeu, Maria José, Delano and Marcelo, for their love and support. Their words of encouragement and caring supported me in the most difficult moments. I take this opportunity to apologize for the physical absence in all times they needed me.

Along the way, life presented me with angels, who will always have a very special place in my heart. Thanks, Melina and Marcela, for being my family in Piracicaba, for the friendship, the countless moments of fun and for the harmony and excellent atmosphere in our home.

Thanks to my angels in Belgium, Lies and Tiago, for coloring my gray days. You were my support and were essential to make the time away from home the most enjoyable as possible. Our laughs, chats, dinners, trips and friendship will always be in my fondest memories.

I also thank my professors of LCE, for the valuable teachings and advices over the years. Thanks to the others staff, for constant support and patience, and to my colleagues of LCE for the company during the studies, lectures and courses. A special thanks to Elizabeth Hashimoto, for the good conversations, advices and help during several occasions.

I thank my friends from CenStat, especially Elasma Milanzi, Edmund Njagi, Chenjerai Mutambanengwe, Sammy Chebon and Chellafe Ensoy, for the constant help. I learned a lot from you! Thank you, my dear friends in Belgium, Renata Yokota, Martin Otava, Reza and Aida, for the warm welcome and friendship. The stay in Belgium for sure would not have been so pleasant without you.

My dear friends in Piracicaba: Rafaela Naves, thanks for welcoming me in the city and for the great friendship that was initiated during those days; Luciana Chaves, thanks for the good laughs, advices and excellent company; Fernando Toledo, thanks for helping me with $\LaTeX$ and genetics; Matheus Nunes, thanks for the pleasurable discussions on statistics, forest engineering and life.

Lastly, I thank my grandmother, Ana de Lourdes, who couldn't see the completion of this work, but she has always been my greatest example of strength, courage and faith.

SUMMARY

RESUMO	9
ABSTRACT	11
SAMENVATTING	13
1 INTRODUCTION	15
References	17
2 HERITABILITY BASED ON FLEXIBLE MODELS FOR HIERARCHICAL AND OVERDISPERSED TIME-TO-EVENT DATA	19
Abstract	19
2.1 Introduction	19
2.2 Motivating Case Study	20
2.3 Background on Heritability	21
2.4 Modeling Framework	22
2.4.1 Generalized Linear Models	22
2.4.2 Overdispersion Models	23
2.4.3 Models With Normal Random Effects	24
2.4.4 Models Combining Overdispersion With Normal Random Effects	24
2.5 Closed-form Derivation of Heritability	25
2.5.1 Precision estimation	27
2.6 Data Analysis	29
2.7 Concluding Remarks	31
References	31
3 QUANTIFYING THE GENETIC CONTRIBUTION TO THE VARIABILITY OF COUNT TRAITS	35
Abstract	35
3.1 Introduction	35
3.2 Motivating Case Study	36
3.3 An Extended Poisson Model to Handle Hierarchical and Overdispersed Data	37
3.4 Derivation of Heritability for Count Data	40
3.5 Data Analysis	42
3.5.1 The Conventional Linear Analysis	42
3.5.2 The Proposed Poisson-type Analysis	42
3.6 Concluding Remarks	45
References	46
4 NEGATIVE VARIANCE COMPONENTS FOR NON-NEGATIVE HIERARCHICAL DATA	49
Abstract	49
4.1 Introduction	49

4.2 Background on Poisson Models	50
4.3 The Case of Random Intercepts and Independent Gamma Variables	52
4.3.1 Variance Component Induced by the Gamma Random Effect	53
4.3.2 Variance Component Induced by the Normal Random Effect	54
4.4 Estimation	54
4.5 A Maize Breeding Study	54
4.6 Concluding Remarks	56
References	57
5 GENERAL CONCLUSIONS AND FUTURE RESEARCH	59
APPENDIX	61

RESUMO

Estratégias de modelagem para dados hierárquicos complexos e com superdispersão em ciências biológicas

Neste trabalho foram estudados os chamados modelos combinados, modelos lineares generalizados mistos com extensão para acomodar superdispersão, no contexto de genética e melhoramento. Esses modelos flexíveis acomodam correlação induzida por agrupamento e superdispersão por meio de dois conjuntos separados de efeitos aleatórios e contêm como casos especiais os modelos lineares generalizados mistos (MLGM) e os modelos de superdispersão comumente conhecidos. Tais modelos são usados na obtenção do coeficiente de herdabilidade para caracteres não Gaussianos. Herdabilidade é um dos vários importantes conceitos que são frequentemente quantificados com o ajuste de um modelo a dados hierárquicos. Ela é usualmente importante no melhoramento vegetal e animal. Conhecer esse atributo é útil para quantificar a magnitude do ganho na população. Para dados em que modelos lineares podem ser usados, esse atributo é convenientemente definido como uma razão de componentes de variância. Os problemas são menos simples para respostas não Gaussianas. O foco aqui é em características do tipo tempo-até-evento e contagem, em que os modelos Weibull-Gama-Normal e Poisson-Gama-Normal são usados. As expressões resultantes são suficientemente simples e atrativas, em particular nos casos especiais, pelo valor prático. As metodologias propostas são ilustradas usando dados de melhoramento animal e vegetal. Além disso, a atenção é voltada à ocorrência de estimativas negativas de componentes de variância no modelo Poisson-Gama-Normal. A ocorrência de componentes de variância negativos em modelos lineares mistos (MLM) tem recebido certa atenção na literatura enquanto quase nenhum trabalho tem sido feito para MLGM. Esse fenômeno pode ser confuso a princípio porque, por definição, variâncias são quantidades não-negativas. Entretanto, este é um fenômeno bem compreendido no contexto de modelagem linear mista, em que a escolha deverá ser feita entre uma interpretação hierárquica ou marginal. Os componentes de variância do modelo combinado para respostas de contagem são estudados teoricamente e o estudo de melhoramento vegetal usado como ilustração confirma que esse fenômeno pode ser comum em pesquisas aplicadas. A atenção também é voltada ao desempenho de diferentes métodos de estimação, porque nem todos aqueles disponíveis são capazes de estender o espaço paramétrico dos componentes de variância. Então, quando há a necessidade de inferência de tais componentes e é esperado que eles sejam negativos, a acurácia do método de estimação não é a única característica a ser considerada.

Palavras-chave: Modelo combinado; Distribuição gama; Modelo linear misto generalizado; Herdabilidade; Variâncias negativas; Distribuição Poisson; Efeito aleatório; Distribuição Weibull

ABSTRACT

Modeling strategies for complex hierarchical and overdispersed data in the life sciences

In this work, we study the so-called combined models, generalized linear mixed models with extension to allow for overdispersion, in the context of genetics and breeding. Such flexible models accommodates cluster-induced correlation and overdispersion through two separate sets of random effects and contain as special cases the generalized linear mixed models (GLMM) on the one hand, and commonly known overdispersion models on the other. We use such models while obtaining heritability coefficients for non-Gaussian characters. Heritability is one of the many important concepts that are often quantified upon fitting a model to hierarchical data. It is often of importance in plant and animal breeding. Knowledge of this attribute is useful to quantify the magnitude of improvement in the population. For data where linear models can be used, this attribute is conveniently defined as a ratio of variance components. Matters are less simple for non-Gaussian outcomes. The focus is on time-to-event and count traits, where the Weibull-Gamma-Normal and Poisson-Gamma-Normal models are used. The resulting expressions are sufficiently simple and appealing, in particular in special cases, to be of practical value. The proposed methodologies are illustrated using data from animal and plant breeding. Furthermore, attention is given to the occurrence of negative estimates of variance components in the Poisson-Gamma-Normal model. The occurrence of negative variance components in linear mixed models (LMM) has received a certain amount of attention in the literature whereas almost no work has been done for GLMM. This phenomenon can be confusing at first sight because, by definition, variances themselves are non-negative quantities. However, this is a well understood phenomenon in the context of linear mixed modeling, where one will have to make a choice between a hierarchical and a marginal view. The variance components of the combined model for count outcomes are studied theoretically and the plant breeding study used as illustration underscores that this phenomenon can be common in applied research. We also call attention to the performance of different estimation methods, because not all available methods are capable of extending the parameter space of the variance components. Then, when there is a need for inference on such components and they are expected to be negative, the accuracy of the method is not the only characteristic to be considered.

Keywords: Combined model; Gamma distribution; Generalized linear mixed model; Heritability; Negative variance components; Poisson distribution; Random effect; Weibull distribution

SAMENVATTING

Modellerings-strategieën voor complexe hiërarchische gegevens met overdispersie, uit de levenswetenschappen

Dit werk is toegewijd aan het zogenaamde gecombineerde model, gesitueerd binnen de familie van de veralgemeend lineaire modellen die ook overdispersie toelaten. De context is genetica en veredeling. Dergelijke modellen laten op flexibele manier correlatie geïnduceerd door clustering toe. Tevens houden ze rekening met overdispersie. Dit gebeurt aan de hand van twee types random effecten. Het veralgemeend lineair gemengd model is er een bijzonder geval van. Hetzelfde geldt voor klassieke overdispersie-modellen. We maken gebruik van dit soort modellen voor het afleiden van erfelijkheidscoëfficiënten voor niet-Gaussiaanse respons. Erfelijkheid is één van de vele belangrijke concepten die vaak worden gekwantificeerd via het fitten van een model aan hiërarchische gegevens. Er wordt groot belang aan gehecht bij het veredeling van plant en dier. Kennis van dit attribuut is belangrijk om de veredeling in kaart te brengen. Wanneer lineaire modellen kunnen gebruikt worden, is een en ander vrij eenvoudig. Immers, de coëfficiënten kunnen dan geschreven worden als verhouding van varianties. Dit verandert wanneer de responsveranderlijke niet normaal verdeeld is. We leggen de klemtoon op respons die de vorm aanneemt van tijden of aantallen. We gebruiken in voorkomend geval het Weibull-Gamma-Normaal of Poisson-Gamma-Normaal model. Ook in dit geval vinden we uitdrukkingen die voldoende eenvoudig zijn en die een aantrekkelijke interpretatie hebben. Dit is a fortiori waar voor een reeks speciale gevallen. De voorgestelde modellen worden geïllustreerd aan de hand van gegevens uit plant- en dierveredeling. Speciale aandacht wordt geschonken aan het voorkomen van negatieve variantiecomponenten in het Poisson-Gamma-Normale model. Het voorkomen van dergelijke negatieve variantiecomponenten in het lineaire gemengde model kreeg enige aandacht in de literatuur, maar hetzelfde geldt niet voor veralgemeend lineaire gemengde modellen. Dit kan verwarring zaaien op het eerste gezicht, omdat varianties als dusdanig niet negatief kunnen zijn. Voor het lineair gemengde model is het fenomeen tamelijk goed begrepen, omdat ook het verschil tussen een hiërarchische en een marginale interpretatie uitgeklaard werd. We bestuderen de variantiecomponenten van het gecombineerde model vanuit theoretisch oogpunt. Het voorbeeld uit de plantveredeling onderstreept dat het fenomeen wel degelijk in de praktijk kan voorkomen. We vestigen de aandacht op verschillen tussen onderscheiden schattingsmethoden. Inderdaad, niet alle methoden laten het zomaar toe van het bereik van variantiecomponenten uit te breiden in het negatieve. Dus, als er nood is aan de mogelijkheid op negatieve variantiecomponenten, moeten we bij de het kiezen van de schattingsmethode niet louter rekening houden met de numerieke precisie.

Staaikwoorden: Erfelijkheid; Gamma verdeling; Gecombineerd model; Negatieve variantiecomponenten; Poisson verdeling; Random effect; Veralgemeend lineair gemengd model; Weibull verdeling

1 INTRODUCTION

Correlated data are common in many areas, such as agriculture, biology, economics and manufacturing. Examples of such data include hierarchical or multilevel data, repeated measurements and longitudinal data. Regarding the last two data types, it is always important to emphasize the difference between them. A design where, for each subject, a characteristic is measured under several experimental conditions is termed a repeated measures study. As a special case, if this characteristic is taken repeatedly over time, the study is termed longitudinal.

The general linear mixed model is a flexible tool for the analysis of correlated data (PINHEIRO; BATES, 2000; VERBEKE; MOLENBERGHS, 2000). Some reasons for its increasing popularity are its flexibility in modeling the within-group correlation and availability of efficient software for fitting it, such as the SAS procedure MIXED (LITTELL et al., 1996), the nlme R package (PINHEIRO et al., 2013) and ASReml (GILMOUR et al., 2009).

Linear mixed models originated in designed experiments (ROBINSON, 1991) and have been widely used in agricultural research. Hierarchical data structures occur frequently in agricultural experimentation and well-known examples are the blocks designed experiments. In plant and animal breeding a family-based structure is considered to evaluate the variability in traits of varieties and offspring. In this context, the heritability coefficient, which is the proportion of the genetic contribution over the total variability in the phenotype, is an important attribute to be calculated. When linear mixed models are used, this measure takes the form of ratios of simple functions of variance components, because the linear relationship between the response and the other terms in the model allows the random terms to be easily separated.

Linear mixed models assume that both the random effects and the errors follow Gaussian distributions and are intended for correlated data in which the response variable is, at least approximately, continuous. Next to continuous outcomes, counts and time-to-event outcomes feature prominently in applied modeling. In agriculture and livestock, and turning to count data, one may be interested, for instance, in number of grains, number of branches of a plant, number of leaves infected by a plant disease, number of offspring of a sire or number of eggs laid. Regarding time to event data, time to flowering, time the seeds take to germinate, length of gestation, lactation period, time to weaning and outcomes related to longevity are common in these areas of research.

One often models such data using generalized linear modeling (GLM) framework (NELDER; WEDDERBURN, 1972; MCCULLAGH; NELDER, 1989), which covers widely used statistical models, such as log-linear models for count and time-to-event data. In this context, the exponential family provides an elegant and encompassing mathematical framework, since it has the normal, Bernoulli/binomial, Poisson and Weibull/exponential as members.

Despite this elegance, a key feature of the GLM framework, the so-called mean-variance

relationship, may be overly restrictive. By this relationship, the variance is a deterministic function of the mean. For example, for count outcomes this relationship is $v(\mu) = \mu^2$. On the other hand, for normally distributed outcomes, the mean and variance are entirely separated parameters.

In certain applications the mean-variance relationship imposed by these models may not be met in a particular set of data. When the data exhibit greater variability than is predicted by this relationship, the phenomenon is referred to as overdispersion, while the opposite case is known as underdispersion. Early work has been developed to formulating models that allow for overdispersion. Overdispersed Poisson data received attention in Breslow (1984) and Lawless (1987). Hinde and Demétrio (1998) provides an overview of approaches for dealing with this phenomenon, placing most emphasis on the binomial and Poisson settings.

In discrete hierarchical data modeling, the so-called generalized linear mixed model (GLMM) (BRESLOW; CLAYTON, 1993; MOLENBERGHS; VERBEKE, 2005) has gained popularity. They are a combination of general exponential family models with normally distributed random effects and part of their popularity is due to the availability of implementations in various software packages.

There are situations where overdispersion and the need for hierarchical modeling occur simultaneously. For such cases, Molenberghs, Verbeke and Demétrio (2007) specified a model for count data and, later, Molenberghs et al. (2010) proposed a broad class of such, where the binary, count, and time-to-event cases were given particular emphasis. The so-called combined models accommodate overdispersion and clustering through two separate sets of random effects and contain as special cases the GLMM on the one hand, and commonly known overdispersion models, such as the negative-binomial, the beta-binomial, and the gamma frailty models, on the other.

In a variety of applied fields, apart from the fixed effects, there is a need for inference on variance components, such as linear mixed models (VERBEKE; MOLENBERGHS, 2000) and generalized linear mixed models (MOLENBERGHS; VERBEKE, 2005). Although variance components are positive by definition, there are plenty of cases in which they may have negative estimates. This phenomenon has received a certain amount of attention in the literature in the context of linear mixed modeling, where one will have to make a choice between a hierarchical and a marginal view. On the other hand, almost no work has been done for generalized linear mixed models (GLMM).

The following chapters present the results of three manuscripts based on modeling strategies for hierarchical and overdispersed data in agricultural sciences, using the aforementioned combined models. The focus of this thesis is on count and time-to-event data.

In Chapter 2 a methodology to calculate heritability for time-to-event traits is proposed. As alluded to in the third paragraph, for data where linear models can be used, the calculation of this attribute is easy. However, matters are less simple for non-Gaussian outcomes. The Weibull combined model (Weibull-Gamma-Normal) and its special case, the exponential combined

model (Exponential-Gamma-Normal) are used to describe the data. These models admit closed-form expressions for the mean, variance, and higher-order moments. As a result, correlation coefficients and variance ratios have explicit expressions too. Then, the derivation of heritability measures is not problematic. The proposed methodology is illustrated using progeny data of Nellore, a Zebu breed.

The focus of Chapter 3 is on count outcomes. Heritability expressions for count traits are derived using the Poisson combined model (Poisson-Gamma-Normal), which combines normal and gamma random effects to capture both correlation among repeated observations as well as overdispersion. The methodology is illustrated using maize data provided by the Plant Breeding Program of Department of Genetics of ESALQ. The results are compared with the conventional, but questionable analysis using linear mixed models. The heritability coefficients obtained in Chapters 2 and 3 are sufficiently simple and appealing, in particular in special cases, to be of practical value.

In Chapter 4 the focus is on the variance components of the Poisson-Gamma-Normal model. The variance components induced by the gamma random effect as well as those induced by the normal random effect are theoretically studied and, using maize data, it is underscored that negative estimates can occur. Such phenomenon is especially common when the variability is low and there is an expected negative intraclass correlation. Furthermore, performances of different estimation methods while estimating the negative variance components are briefly discussed.

Finally, general conclusions and recommendations for future research are presented in last chapter.

References

BRESLOW, N. E. Extra-poisson variation in log-linear models. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, London, v. 33, n. 1, p. 38-44, 1984.

BRESLOW, N. E.; CLAYTON, D. G. Approximate inference in generalized linear mixed models. **Journal of the American Statistical Association**, Boston, v. 88, n. 421, p. 9-25, 1993.

GILMOUR, A. R.; GOGEL, B. J.; CULLIS, B. R.; THOMPSON, R. **ASReml User Guide Release 3.0**. Hemel Hempstead, HP1 1ES, UK: VSN International Ltd, 2009. Disponível em: Www.vsnl.co.uk. Acesso em: 28 maio 2014.

HINDE, J.; DEMÉTRIO, C. G. B. Overdispersion: Models and estimation. **Computational Statistics & Data Analysis**, Amsterdam, v. 27, n. 2, p. 151-170, 1998.

LAWLESS, J. F. Negative binomial and mixed poisson regression. **The Canadian Journal of Statistics**, Ottawa, v. 15, n. 3, p. 209-225, 1987.

LITTELL, R. C.; MILLIKEN, G. A.; STROUP, W. W.; WOLFINGER, R. D. **SAS System for Mixed Models**. Cary, NC: SAS Institute Inc., 1996.

MCCULLAGH, P.; NELDER, J. A. **Generalized linear models**. London: Chapman & Hall, 1989. 532 p.

MOLENBERGHS, G.; VERBEKE, G. **Models for discrete longitudinal data**. New York: Springer, 2005. 683 p.

MOLENBERGHS, G.; VERBEKE, G.; DEMÉTRIO, C. G. B. An extended random-effects approach to modeling repeated, overdispersed count data. **Lifetime Data Analysis**, Hingham, v. 13, n. 4, p. 513-531, 2007.

MOLENBERGHS, G.; VERBEKE, G.; DEMÉTRIO, C. G. B.; VIEIRA, A. A family of generalized linear models for repeated measures with normal and conjugate random effects. **Statistical Science**, Hayward, v. 25, n. 3, p. 325-347, 2010.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. **Journal of the Royal Statistical Society: Series A**, London, v. 135, n. 3, p. 370-384, 1972.

PINHEIRO, J.; BATES, D.; DEBROY, S.; SARKAR, D.; R Core Team. **nlme**: Linear and Nonlinear Mixed Effects Models. [S.l.], 2013. R package version 3.1-113.

PINHEIRO, J. C.; BATES, D. M. **Mixed-Effects Models in S and S-PLUS**. New York: Springer-Verlag, 2000. 528 p.

ROBINSON, G. K. That blup is a good thing: The estimation of random effects. **Statistical Science**, Hayward, v. 6, n. 1, p. 15-32, 1991.

VERBEKE, G.; MOLENBERGHS, G. **Linear mixed models for longitudinal data**. New York: Springer-Verlag, 2000. 568 p.

2 HERITABILITY BASED ON FLEXIBLE MODELS FOR HIERARCHICAL AND OVERDISPERSED TIME-TO-EVENT DATA

Abstract

Heritability is one of the many important concepts that are often quantified upon fitting a model to hierarchical data. Other quantities of this type, from different areas of application, include inter-rater agreement, (test-retest) reliability, etc. When based on Gaussian data, this type of quantities conveniently takes the form of ratios of linear combinations of variance components. Matters are less simple for alternative outcome types. The focus here is on time-to-event outcomes, where a so-called combined model is used to describe the data. This model combines a Weibull outcome distribution with normal as well as gamma random effects; it flexibly allows for both correlation due to data hierarchies as well as overdispersion. Furthermore, because the model admits closed-form expressions for the mean, variances, higher moments, and even the joint marginal distribution, the derivation of heritability measures is not problematic. Beneficially, the resulting expressions do not depend on the fixed-effects structure (covariates) and hence easily interpretable constant coefficients emanate. Expressions are derived and illustrated using progeny data of Nellore, a Zebu breed.

Keywords: Combined model; Gamma distribution; Generalized linear mixed model; Random effect; Weibull distribution

2.1 Introduction

In animal and plant breeding, as in human biological applications, the determination of so-called heritability is often of importance. Heritability refers to the proportion of the total phenotypic variation that can be explained by the genetic component. It is used to quantify the potential response to selection and, for example, many practical agricultural decisions about a breeding procedure depend on its magnitude. Its determination is routinely based on hierarchical data, of a progeny- or family-based nature. When the outcomes are, perhaps approximately, normally distributed, linear (mixed) models are frequently used. Heritability measures then take the form of ratios of simple functions of variance components. Indeed, linear mixed models (VERBEKE; MOLENBERGHS, 2000) have been widely used to estimate, and separate from each other, the genetic and environmental effects by considering these factors as random terms in the model. The linear relationship between the response and the other terms in the model allows that heritability be obtained easily. The ASReml (GILMOUR et al., 2009) statistical package has been widely used in agricultural and livestock studies to estimate variance components and heritability (ALBUQUERQUE; MEYER, 2001; CARVALHEIRO et al., 2006; NEVES et al., 2012). The package can easily handle quite large data sets with complex variance models, required for ‘animal models.’ It can also be used to fit the linear mixed models under investigation.

Apart from Gaussian outcomes, time-to-event data are common too in applied modeling in

several areas of science. In livestock, for example, one may have interest in longevity, length of gestation, number of days from the first parity to the culling day, lactation period, time to weaning, and others. It is common to model such data using generalized linear modeling and survival analysis. In applied research, it has been clear for time-to-event data that the often mathematically elegant but practically restrictive mean-variance relationship imposed by this class of models may not be met. Thus, many papers were devoted to formulate models for dealing with these phenomena called under- or overdispersion.

As alluded to in the opening paragraph, studies of the type considered imply the collection of hierarchical data. Apart from the aforementioned linear mixed models, for non-Gaussian outcomes, the so-called generalized linear mixed models (GLMM) has gained popularity (MOLENBERGHS; VERBEKE, 2005).

There are situations where overdispersion and the need for hierarchical modeling occur simultaneously. Combining ideas from the overdispersion models and models with normal random effects, Molenberghs et al. (2010) proposed a broad class of generalized linear models that accommodate overdispersion and clustering through two separate sets of random effects, termed the combined model. This class of models contains as special cases the GLMM on the one hand, and traditional overdispersion models such as the negative-binomial and beta-binomial models on the other.

In this paper, we use such models for handling overdispersion and correlated data, while obtaining heritability coefficients based on time-to-event traits. The components of variability present in the model allow the separation of genetic from residual components. Importantly, the resulting coefficient will be seen to be free of fixed-effects, leading to simple, easily interpretable quantities.

The proposed methodology will be illustrated using data from a study in animal sciences. The data refer to time that progenies of Nellore sires, a Zebu breed, take to reach a pre-specified weight. This time-to-event trait is of interest in breeding procedures such that the aim is a shorter time. We derive the proportion of its variability that is due to genetic effects, i.e., heritability.

The paper is organized as follows. In Section 2.2, the motivating case is described with the analysis reported in Section 2.6. In Section 2.3, we present the main concepts of heritability. A brief review of useful models, including the combined model for time-to-event data, is the subject of Section 2.4. We use this combined model to obtain the heritability coefficient for time-to-event traits and the expressions are presented in Section 2.5.

2.2 Motivating Case Study

The data considered in this work were first analyzed in Giolo and Demétrio (2011). They refer to records from progeny of Nellore, a Zebu breed from India that has become predominant in Brazil. The original database consists of 3611 progenies of 24 Nellore sires and 3116 dams born during spring in a single herd between 1996 and 1997. About 16% of the dams had two

Table 2.1 – Descriptive statistics for estimated time (months) per sires that their progeny take to achieve the specified weight of 160 kg

Mean	St. dev.	<i>N</i>	Max.	Min.
19.58	0.93	16	20.67	16.81
18.77	0.93	79	21.15	15.83
18.57	0.95	198	20.82	16.43
⋮	⋮	⋮	⋮	⋮
17.51	0.97	32	19.37	15.96
17.72	0.90	150	20.16	15.14
17.89	0.90	193	20.03	15.62

progenies. For simplicity, we randomly selected one of the offspring of each of these dams. As a result, the data set used here consisted of 3116 progenies and the number of progenies per sire varied from 16 to 269. Evidently, a sire random effect will be needed.

All progenies were followed up from birth to approximately 2.5 years after they were born. Over this period, the weight of each progeny was taken six times at intervals of approximately 3–5 months, where the first weight was taken at birth. In addition to the weight, information about sex, reproduction (natural or artificial), progeny birth year, and age of the dam at progeny birth were recorded for each progeny.

Giolo and Demétrio (2011) compared sires based on length in days that their progenies need to gain 160 kg, a commercially specified weight gain employed in Brazil. Given that the weight is taken only periodically, the exact time that each progeny takes to gain the weight of interest was not precisely known. Thus, a logistic growth curve model for each progeny was fitted to estimate the exact response time. There are no censored data because all progenies have gained the targeted weight over the follow-up period.

In line with these authors, the estimated time to reach 160 kg is the response time of interest. To avoid (computational) scale problems during the model fitting, we use the months unit rather than the days one; evidently, this has no implications from a statistical view-point. Table 2.1 shows descriptive statistics for the estimated time, in months, that progenies took to gain 160 kg. For illustrative purposes we take the three sires whose progenies took the longest time and the three sires of which the progenies took the shortest time to gain this weight. More details about the data set and the estimation process are described in Giolo and Demétrio (2011).

2.3 Background on Heritability

Heritability is relatively easy to examine for Gaussian traits. Generally, the heritability of a characteristic quantifies the proportion of the total variation in a phenotypic characteristic that can be explained by genotypic effects (FALCONER; MACKAY, 1996; VISSCHER et al., 2008).

Based on a linear model, the total variation of observed phenotypes (*P*) of a trait can be decomposed into the sum of unobserved genotypic factors (*G*) and unobserved environmental

factors (E), with variances σ_G^2 and σ_E^2 . Furthermore the genetic variance (σ_G^2) can be partitioned into the variance of additive genetic effects (breeding values; σ_A^2) and the variance of non-additive genetic effects of dominance (σ_D^2) and of epistatic interactions (σ_I^2).

From these decompositions and for data where linear models can be used, the heritability is simply defined as a ratio of variances. In a broad sense, it is expressed as $H^2 = \sigma_G^2 / \sigma_P^2$ and in a narrow (strict) sense it is defined as $h^2 = \sigma_A^2 / \sigma_P^2$. This partitioning of the phenotypic variance assumes the absence of genotypic by environmental covariance ($\sigma_{G,E}$) and the interaction between genotype and environment ($G \times E$).

Traditionally, heritability was estimated from simple and balanced designs using parent-offspring regressions and variance components from an analysis of variance, for instance. But when the phenotypic measures are available for individuals with different relationships or when the design is unbalanced, the estimates of the variance components are conveniently obtained from mixed models. Unsurprisingly therefore, linear mixed models have been widely used by breeders. The linear structure of these models allows that the phenotypic variance be decomposed in the sum of genetic and environmental random components and a residual component.

Specifically in livestock studies the most common choice is the ‘animal model,’ which includes the additive genetic effect of animals as a random effect and incorporates all information on animals’ pedigrees using a relationship matrix.

However, when the trait of interest is not normally distributed and/or it does not follow a linear model, the genetic and environmental random terms are no longer easily separated. In time-to-event responses, for example, the relationship between the observed responses and the random effects is not linear and heritability coefficient cannot be easily obtained from conventional survival models. This situation will be addressed in subsequent sections.

2.4 Modeling Framework

In Section 2.4.1, we will briefly review the exponential family and generalized linear modeling (MCCULLAGH; NELDER, 1989). Sections 2.4.2 and 2.4.3 are devoted to reviews of models for overdispersion and generalized linear mixed models (GLMM) respectively (MOLENBERGHS; VERBEKE, 2005). In Section 2.4.4 the model combining correlated data and overdispersion for time-to-event data is presented (MOLENBERGHS et al., 2010).

2.4.1 Generalized Linear Models

A random variable Y follows a distribution of the exponential family if its density can be written as

$$f(y) = f(y|\theta, \phi) = \exp\{\phi^{-1}[y\theta - \psi(\theta)] + c(y, \phi)\}, \quad (2.1)$$

for unknown parameters θ and ϕ , and for known functions $\psi(\cdot)$ and $c(\cdot, \cdot)$. Usually, θ and ϕ are termed ‘natural (canonical) parameter’ and ‘scale parameter,’ respectively. The first two

moments follow from $\psi(\cdot)$ (MOLENBERGHS; VERBEKE, 2005) as $E(Y) = \mu = \psi'(\theta)$ and $\text{Var}(Y) = \sigma^2 = \phi\psi''(\theta)$. In general, the mean and variance are related through $\sigma^2 = \phi\psi''[\psi'^{-1}(\mu)] = \phi v(\mu)$, with $v(\mu)$ the so-called variance function.

In some circumstances, the variance function $v(\mu)$ can be chosen in accordance with a particular member of the exponential family. If not, the parameters cannot be estimated using maximum likelihood principles. Instead, a set of estimating equations needs to be specified, the solution of which is referred to as the quasi-likelihood estimates (MCCULLAGH; NELDER, 1989; MOLENBERGHS; VERBEKE, 2005).

In a regression context, one wishes to explain variability between N independent outcomes $Y_1, \dots, Y_N$ based on $x_1, \dots, x_N$, p -dimensional vectors of covariate values. This leads to so-called generalized linear models. It is assumed that all Y_i have densities $f(y_i|\theta_i, \phi)$ that belong to the exponential family, but a different parameter is allowed per observation. Specification of the model is completed by modeling the means μ_i : $\mu_i = h(\eta_i) = h(\mathbf{x}_i'\boldsymbol{\xi})$, for a known function $h(\cdot)$ and with $\boldsymbol{\xi}$, a vector of p fixed and unknown regression coefficients. Usually, $h^{-1}(\cdot)$ is called the link function. The natural link function is given by $h(\cdot) = \psi'(\cdot)$.

2.4.2 Overdispersion Models

In certain applications of standard generalized models it is often found that the data exhibit overdispersion, i.e., the variability is greater than that prescribed by the mean-variance relationship. When this phenomenon is not taken into account severe estimation problems in the variance-covariance matrix can occur and/or misleading inferences may ensue. A number of models have been proposed for handling overdispersion. Hinde and Demétrio (1998) provide a general treatment of overdispersion; the Poisson case received particular attention by Breslow (1984) and Lawless (1987).

A straightforward step is to allow the overdispersion parameter $\phi \neq 1$, so that $\text{Var}(Y) = \phi\mu$. An elegant mechanism is through a two-stage approach for the response, by assuming that the model parameter follows a particular distribution.

The frailty model was introduced in survival analysis by Vaupel et al. (1979) as a way to allow unobserved heterogeneity. The model is an extension of the proportional hazards model in which the frailty term acts multiplicatively on the baseline hazard and captures the individual heterogeneity that refers to unobserved risk factors. It is assumed that this frailty term follows a parametric distribution; the most commonly used ones are gamma and log-normal. The gamma distribution has been preferred for its flexibility and algebraic convenience, given that it is possible to obtain explicit expressions for the marginal functions when the density function is integrated with respect to the frailty terms. A detailed treatment is given in Duchateau and Janssen (2008).

In clustered survival data, a correlated frailty approach has been used in several studies. In this model, the frailty is allowed to be individual-specific, but unlike in the univariate case there is no assumption of independence and one individual's frailty can be associated with the

frailty of another individual who is genetically related to the first. Correlated frailty models have received attention in animal breeding studies for analyzing traits associated with productive life of livestock (SCHUKKEN et al., 2010; GIOLO; DEMÉTRIO, 2011).

2.4.3 Models With Normal Random Effects

The generalized linear mixed model (GLMM) (BRESLOW; CLAYTON, 1993; MOLENBERGHS; VERBEKE, 2005) arguably is the most frequently used random-effects model for non-Gaussian outcomes. It is a straightforward extension of the generalized linear model for independent data (2.4.1) to the context of clustered measurements, on the one hand, and of the linear mixed model, on the other. To enhance its use, there is a wide range of software tools available for fitting the models.

Conditionally on random effects $\mathbf{b}_i$, usually assumed to be $N(\mathbf{0}, D)$ distributed, the GLMM assumes that the elements Y_{ij} of $\mathbf{Y}_i$ are independent with densities of the form

$$f_i(y_{ij}) = f_i(y_{ij}|\theta_{ij}, \phi) = \exp\{\phi^{-1}[y_{ij}\theta_{ij} - \psi(\theta_{ij})] + c(y_{ij}, \phi)\}, \quad (2.2)$$

where $h^{-1}(\mu_{ij}) = h^{-1}[E(Y_{ij}|\mathbf{b}_i)] = x'_{ij}\boldsymbol{\xi} + z'_{ij}\mathbf{b}_i$, x_{ij} and z_{ij} are p -dimensional and q -dimensional vectors of known covariate values, $\boldsymbol{\xi}$ is a p -dimensional vector of unknown fixed regression coefficients and ϕ is the scale parameter.

When normality of the random effects is assumed, the mean vector and variance-covariance matrix of $\mathbf{Y}_i$ can be derived without too much difficulty. The expressions of the Weibull-Normal model are presented in Appendix A. This kind of model is less common for survival data, where the frailty models, described in the previous section, are more standard.

2.4.4 Models Combining Overdispersion With Normal Random Effects

Combining ideas from overdispersion models and models with normal random effects, Molenberghs et al. (2007) specified a model for count data. Later, Molenberghs et al. (2010) proposed a broad class of such, where the binary, count, and time-to-event cases were given particular emphasis. These models accommodate overdispersion and clustering through two separate sets of random effects and produce models with only random effects and models with only overdispersion as special cases.

Time-to-event combined models allow for closed-form expressions for the mean vector and variance-covariance matrix. As highlighted by Molenberghs et al. (2010), the derivation of such closed forms has important implications because they allow, for example, further derivation of explicit forms for correlation functions. This aspect was covered by Vangeneugden et al. (2011) for Poisson type models for count data, with normal and gamma random effects.

The Weibull model for repeated measures, with both gamma and normal random effects,

can be expressed as

$$f(\mathbf{y}_i | \boldsymbol{\theta}_i, \mathbf{b}_i) = \prod_{j=1}^{n_i} \rho \theta_{ij} y_{ij}^{\rho-1} e^{x'_{ij} \boldsymbol{\xi} + z'_{ij} \mathbf{b}_i} e^{-y_{ij}^{\rho} \theta_{ij} e^{x'_{ij} \boldsymbol{\xi} + z'_{ij} \mathbf{b}_i}}, \quad (2.3)$$

$$f(\boldsymbol{\theta}_i) = \prod_{j=1}^{n_i} \frac{1}{\beta_j^{\alpha_j} \Gamma(\alpha_j)} \theta_{ij}^{\alpha_j-1} e^{-\theta_{ij}/\beta_j}, \quad (2.4)$$

$$f(\mathbf{b}_i) = \frac{1}{(2\pi)^{q/2} |D|^{1/2}} e^{-(1/2) \mathbf{b}_i' D^{-1} \mathbf{b}_i}, \quad (2.5)$$

where ρ is the Weibull shape parameter, $\boldsymbol{\xi}$ is a vector of fixed effects parameters, $\mathbf{b}_i$ is a vector of normal random effects and θ_{ij} are the gamma random effects. These effects are assumed independent while the correlation between the repeated measures is induced by normal random effects.

Setting $\rho = 1$ leads to the special case of an exponential time-to event distribution. It is also evident that the classical gamma frailty model (no normal random effects) and the Weibull-based GLMM (no gamma random effects) follow as special cases.

The above expressions are derived for a two-parameter gamma density and it is customary to set $\alpha_j \beta_j = 1$, for reasons of identifiability, although sometimes alternative restrictions are chosen. Closed-forms expressions for the marginal density, means, variances, and covariances were derived by Molenberghs et al. (2010). The ones needed for our purposes, i.e., to derive closed-form heritability coefficients, are presented in Appendix A.

2.5 Closed-form Derivation of Heritability

We here present expressions for heritability in the Weibull and exponential cases by using the combined model, reviewed in the previous section. Details on calculations, as well as a further generalization, can be found in Appendix B.

Consider the combined model for overdispersion and correlated data presented in (2.3)–(2.5). Molenberghs et al. (2010) showed that its variance equals (1). Assuming $\beta_j = 1/\alpha_j$, the contribution from overdispersion over the total variability is

$$\zeta_{W_{ij}} = \frac{\Gamma(\alpha_j) \Gamma\left(\alpha_j - \frac{2}{\rho}\right) \Gamma\left(\frac{2}{\rho} + 1\right) - \Gamma\left(\alpha_j - \frac{1}{\rho}\right)^2 \Gamma\left(\frac{1}{\rho} + 1\right)^2}{\exp\left(\frac{1}{\rho^2} z'_{ij} D z_{ij}\right) v(\alpha_j, \rho, D, \mathbf{z}_{ij})}, \quad (2.6)$$

with

$$\begin{aligned} v(\alpha_j, \rho, D, \mathbf{z}_{ij}) &= \left[\Gamma(\alpha_j) \Gamma\left(\alpha_j - \frac{2}{\rho}\right) \Gamma\left(\frac{2}{\rho} + 1\right) \exp\left(\frac{1}{\rho^2} z'_{ij} D z_{ij}\right) \right. \\ &\quad \left. - \Gamma\left(\alpha_j - \frac{1}{\rho}\right)^2 \Gamma\left(\frac{1}{\rho} + 1\right)^2 \right]. \end{aligned}$$

The ratio in (2.6) places the variance of the combined model, considering $D = 0$, in the numerator and the full variance of the combined model in the denominator. The heritability coefficient is obtained by:

$$H_{W_{ij}}^2 = 1 - \zeta_{W_{ij}}, \quad (2.7)$$

that is the variability not related to the overdispersion or the proportion in the total variability due to the normal random term, usually the genetic effect.

Very importantly, the above expressions do not depend on the fixed-effects structure. In other words, the only subject-specific influence comes from the random-effects design z'_{ij} . This creates opportunities for constant heritability coefficients, or at least constant over large sub-populations. As an evident example, there can be a dependence on the degree of relationship.

Setting $\rho = 1$ leads to the special case of an exponential time-to-event distribution. For this case the non-genetic contribution is

$$\zeta_{E_{ij}} = \frac{\alpha_j}{\exp(z'_{ij} D z_{ij}) [2(\alpha_j - 1) \exp(z'_{ij} D z_{ij}) - (\alpha_j - 2)]}, \quad (2.8)$$

and the heritability coefficient is obtained by

$$H_{E_{ij}}^2 = 1 - \zeta_{E_{ij}}. \quad (2.9)$$

Specific cases arise when there is only one variance component, d , in D . In this case, $\zeta_{W_{ij}}$ reduces to

$$\zeta_{W_j} = \frac{\Gamma(\alpha_j) \Gamma\left(\alpha_j - \frac{2}{\rho}\right) \Gamma\left(\frac{2}{\rho} + 1\right) - \Gamma\left(\alpha_j - \frac{1}{\rho}\right)^2 \Gamma\left(\frac{1}{\rho} + 1\right)^2}{\exp\left(\frac{1}{\rho^2} d\right) v(\alpha_j, \rho, d)}, \quad (2.10)$$

with

$$v(\alpha_j, \rho, d) = \left[\Gamma(\alpha_j) \Gamma\left(\alpha_j - \frac{2}{\rho}\right) \Gamma\left(\frac{2}{\rho} + 1\right) \exp\left(\frac{1}{\rho^2} d\right) - \Gamma\left(\alpha_j - \frac{1}{\rho}\right)^2 \Gamma\left(\frac{1}{\rho} + 1\right)^2 \right]$$

and $\zeta_{E_{ij}}$ reduces to

$$\zeta_{E_j} = \frac{\alpha_j}{\exp(d) [2(\alpha_j - 1) \exp(d) - (\alpha_j - 2)]}. \quad (2.11)$$

As the Weibull-Normal model is a special case of the combined model, we can also obtain the heritability based on this model. For this, we derived the variance expression for this specific case, that is presented in Appendix A. The non-genetic contribution over the total variability, considering the Weibull-Normal model, is:

$$\zeta_{WN_{ij}} = \frac{\Gamma\left(\frac{2}{\rho} + 1\right) - \Gamma\left(\frac{1}{\rho} + 1\right)^2}{\exp\left(\frac{1}{\rho^2} z'_{ij} D z_{ij}\right) \left[\Gamma\left(\frac{2}{\rho} + 1\right) \exp\left(\frac{1}{\rho^2} z'_{ij} D z_{ij}\right) - \Gamma\left(\frac{1}{\rho} + 1\right)^2 \right]} \quad (2.12)$$

and, considering the Exponential-Normal model, this contribution is:

$$\zeta_{EN_{ij}} = \frac{1}{\exp(z'_{ij} D z_{ij}) [2 \exp(z'_{ij} D z_{ij}) - 1]}. \quad (2.13)$$

2.5.1 Precision estimation

Precision estimates for heritability can be obtained using the delta method (WELSH, 1996). Since $H^2 = 1 - \zeta$, $\text{Var}(H^2) = \text{Var}(\zeta)$. We consider $\zeta = \zeta_N / \zeta_D$, apply the delta method first to numerator and denominator, and then to the ratio. Assume W is the variance-covariance matrix of $\begin{pmatrix} \zeta_N \\ \zeta_D \end{pmatrix}$, then $\text{Var}(\zeta) \cong T'WT$, where

$$T = \frac{\partial \zeta}{\partial (\zeta_N, \zeta_D)} = \begin{pmatrix} 1/\zeta_D \\ -\zeta_N/\zeta_D^2 \end{pmatrix}.$$

To estimate W , we also consider the delta method. At this stage, let ϕ be the parameter vector relevant for ζ_N and ζ_D and V be its variance-covariance matrix. Then, $W \cong S'VS$, with $S = \frac{\partial (\zeta_N, \zeta_D)}{\partial \phi}$.

The V matrix can be easily obtained from available statistical software and the S matrix structure depends on the model used to calculate the heritability. Considering the Weibull-Gamma-Normal model, i.e., the combined model, and for the special case of one variance component in D , the S matrix is

$$\begin{pmatrix} \frac{\partial \zeta_N}{\partial d} & \frac{\partial \zeta_D}{\partial d} \\ \frac{\partial \zeta_N}{\partial \rho} & \frac{\partial \zeta_D}{\partial \rho} \\ \frac{\partial \zeta_N}{\partial \alpha} & \frac{\partial \zeta_D}{\partial \alpha} \end{pmatrix},$$

where

$$\begin{aligned}
\frac{\partial \zeta_N}{\partial d} &= 0, \\
\frac{\partial \zeta_N}{\partial \rho} &= \frac{2}{\rho^2} \Gamma(\alpha) \Gamma\left(\alpha - \frac{2}{\rho}\right) \Gamma\left(\frac{2}{\rho} + 1\right) \left[\psi\left(\alpha - \frac{2}{\rho}\right) - \psi\left(\frac{2}{\rho} + 1\right) \right] \\
&\quad + \frac{2}{\rho^2} \Gamma\left(\alpha - \frac{1}{\rho}\right)^2 \Gamma\left(\frac{1}{\rho} + 1\right)^2 \left[-\psi\left(\alpha - \frac{1}{\rho}\right) + \psi\left(\frac{1}{\rho} + 1\right) \right], \\
\frac{\partial \zeta_N}{\partial \alpha} &= \Gamma(\alpha) \Gamma\left(\alpha - \frac{2}{\rho}\right) \Gamma\left(\frac{2}{\rho} + 1\right) \left[\psi(\alpha) + \psi\left(\alpha - \frac{2}{\rho}\right) \right] \\
&\quad - 2\Gamma\left(\alpha - \frac{1}{\rho}\right)^2 \times \psi\left(\alpha - \frac{1}{\rho}\right) \Gamma\left(\frac{1}{\rho} + 1\right)^2, \\
\frac{\partial \zeta_D}{\partial d} &= \frac{e^{d/\rho^2}}{\rho^2} \left[\Gamma(\alpha) \Gamma\left(\alpha - \frac{2}{\rho}\right) \Gamma\left(\frac{2}{\rho} + 1\right) 2e^{d/\rho^2} - \Gamma\left(\alpha - \frac{1}{\rho}\right)^2 \Gamma\left(\frac{1}{\rho} + 1\right)^2 \right], \\
\frac{\partial \zeta_D}{\partial \rho} &= e^{2d/\rho^2} \Gamma(\alpha) \Gamma\left(\alpha - \frac{2}{\rho}\right) \Gamma\left(\frac{2}{\rho} + 1\right) \left[-\frac{4d}{\rho^3} + \frac{2}{\rho^2} \psi\left(\alpha - \frac{2}{\rho}\right) - \frac{2}{\rho^2} \psi\left(\frac{2}{\rho} + 1\right) \right] \\
&\quad - e^{d/\rho^2} \Gamma\left(\alpha - \frac{1}{\rho}\right)^2 \Gamma\left(\frac{1}{\rho} + 1\right)^2 \left[-\frac{2d}{\rho^3} + \frac{1}{\rho^2} \psi\left(\alpha - \frac{1}{\rho}\right) - \frac{1}{\rho^2} \psi\left(\frac{1}{\rho} + 1\right) \right], \\
\frac{\partial \zeta_D}{\partial \alpha} &= e^{2d/\rho^2} \Gamma(\alpha) \Gamma\left(\alpha - \frac{2}{\rho}\right) \Gamma\left(\frac{2}{\rho} + 1\right) \left[\psi(\alpha) + \frac{2}{\rho^2} \psi\left(\alpha - \frac{2}{\rho}\right) \right] \\
&\quad - 2e^{d/\rho^2} \Gamma\left(\alpha - \frac{1}{\rho}\right)^2 \psi\left(\alpha - \frac{1}{\rho}\right) \Gamma\left(\frac{1}{\rho} + 1\right)^2.
\end{aligned}$$

For the Weibull-Normal case, the S matrix reduces to

$$\begin{pmatrix} \frac{\partial \zeta_N}{\partial d} & \frac{\partial \zeta_D}{\partial d} \\ \frac{\partial \zeta_N}{\partial \rho} & \frac{\partial \zeta_D}{\partial \rho} \end{pmatrix},$$

where

$$\begin{aligned}
\frac{\partial \zeta_N}{\partial d} &= 0, \\
\frac{\partial \zeta_N}{\partial \rho} &= \frac{2}{\rho^2} \left[-\Gamma\left(\frac{2}{\rho} + 1\right) \psi\left(\frac{2}{\rho} + 1\right) + \Gamma\left(\frac{1}{\rho} + 1\right)^2 \psi\left(\frac{1}{\rho} + 1\right) \right], \\
\frac{\partial \zeta_D}{\partial d} &= \frac{1}{\rho^2} e^{d/\rho^2} \left[2\Gamma\left(\frac{2}{\rho} + 1\right) e^{d/\rho^2} - \Gamma\left(\frac{1}{\rho} + 1\right)^2 \right], \\
\frac{\partial \zeta_D}{\partial \rho} &= -\frac{4d}{\rho^3} e^{2d/\rho^2} \Gamma\left(\frac{2}{\rho} + 1\right) - \frac{2}{\rho^2} e^{2d/\rho^2} \Gamma\left(\frac{2}{\rho} + 1\right) \psi\left(\frac{2}{\rho} + 1\right) \\
&\quad + \frac{2}{\rho^3} e^{d/\rho^2} \Gamma\left(\frac{1}{\rho} + 1\right)^2 + \frac{2}{\rho^2} e^{d/\rho^2} \Gamma\left(\frac{1}{\rho} + 1\right)^2 \psi\left(\frac{1}{\rho} + 1\right).
\end{aligned}$$

In the above, the function $\psi(\cdot)$ in the derivatives above refers to the digamma function and the exponential cases follow by setting $\rho = 1$.

2.6 Data Analysis

Giolo and Demétrio (2011) analyzed the data introduced in Section 2.2 using correlated frailty models with the baseline hazard function completely unspecified (Cox model). These authors highlighted that for the mixed-effects Cox model it is not possible to obtain direct heritability estimates since there is no random error variance component. We will re-analyze the data but considering exponential and Weibull models and a predictor of the form:

$$\kappa_{ij} = \xi_0 + b_i + \xi_1 S_i + \xi_2 P_i,$$

where S_i is an indicator for progeny sex (1: male; 0: female), P_i is an indicator for progeny birth year (1: 1996; 0: 1997), and $b_i \sim N(0, d)$.

We consider special cases: (1) the purely exponential model (E--), (2) the purely Weibull model (W--), (3) the Weibull-Gamma model (WG-), (4) the Exponential-Normal model (E-N), (5) the Weibull-Normal model (W-N), and (6) the Weibull-Gamma-Normal model (WGN). Results from fitting all six models are displayed in Table 2.2. Clearly, the goodness-of-fit of the Weibull models is much higher than for the exponential models. This is not surprising, giving the high value of the estimated shape parameter $\hat{\rho}$. Fitting the Exponential-Gamma and Exponential combined models was not possible. This is not a total surprise. Molenberghs and Verbeke (2011) showed that the (WGN) model, and its sub-models, has a finite number of finite moments only. Precisely how many depends on the values of α and ρ : the order k of the moment should be $k < \rho\alpha$ for it to be finite. Given that the estimated ρ is high for the Weibull models, and the α 's are in the order of magnitude of unity, there is no problem for the Weibull-based models considered here. However, setting $\rho = 1$ to obtain exponential versions, and assuming that α would be in the same order of magnitude, might imply that, for example, variances and perhaps means are not finite, underscoring difficulty of fit.

Turning again to the Weibull models, both the (WG-) and the (W-N) models are improvements, in terms of likelihood, relative to the (W--) model. But the strongest improvement in fit occurs when considering the combined (WGN) model. As highlighted by Molenberghs et al. (2010), this strongly affects the point and precision estimates of key quantities such as slope difference and slope ratio. In our case the differences between the fixed effects coefficients estimates of the (WGN) and those of the (W--) and (W-N) models can be clearly noticed. It is not uncommon to consider the (W-N) model, accommodating the correlation between progenies of the same sire, but according to the fit statistics the (WGN) is more appropriate. Note that, while the point estimates for the Weibull shape parameters are high, the corresponding standard errors are small, providing further evidence for proper convergence. Further, note that for the (E-N) the variance of the normal random effect d is estimated on the boundary of the parameter space, effectively removing the parameter from the model. This explains why the (E--) and (E-N) fits are identical. A similar phenomenon does not occur in the Weibull cases.

Table 2.2 – Nelore study. Parameter estimates and standard errors for the regression coefficients in (1) the purely Exponential model, (2) the purely Weibull model, (3) the Weibull-gamma model, (4) the Exponential-normal model, (5) the Weibull-normal model, and (6) the Weibull combined model. Estimation was done by maximum likelihood using numerical integration over the normal effect, if present

Effect	Par.	Estimate (standard error)		
		(E--)	(W--)	(WG-)
Intercept	ξ_0	-2.92 (0.03)	-13.72 (0.24)	-107.19 (2.89)
Sex	ξ_1	0.08 (0.04)	1.11 (0.04)	2.70 (0.10)
Prog. birth	ξ_2	-0.02 (0.04)	0.11 (0.04)	-0.65 (0.06)
Weibull shape	ρ		4.53 (0.08)	36.68 (1.00)
Gamma param.	α			1.53 (0.13)
-2log-likelihood		24314.0	15495.0	7230.9
AIC		24320.0	15503.0	7240.9
Effect	Par.	(E-N)	(W-N)	(WGN)
Intercept	ξ_0	-2.92 (0.03)	-75.01 (0.99)	-110.78 (3.05)
Sex	ξ_1	0.08 (0.04)	1.87 (0.05)	2.81 (0.10)
Prog. birth	ξ_2	-0.02 (0.04)	-0.27 (0.11)	-0.65 (0.17)
Weibull shape	ρ		25.50 (0.33)	37.88 (1.05)
Gamma param.	α			1.52 (0.14)
Var. of progenies	d	6.33×10^{-14}	0.09 (0.03)	0.21 (0.08)
-2log-likelihood		24314.0	7399.0	7100.6
AIC		24322.0	7409.0	7112.6

Based on the model fits, we can now calculate heritability coefficients. Using the one variance component in the random intercept for the sire effect, we apply expression (2.7) using (2.10). Note that throughout Section 2.5, the Gamma distribution parameter α , is represented with an index for the individual j . That said, here and in several other applications α is kept constant. Based on the estimates for the (WGN), as laid out in Table 2.2, we obtain $\zeta = 0.9230$ and:

$$H^2 = 1 - 0.9230 = 0.077(\text{s.e. } 0.7943).$$

The low level of heritability is entirely natural for this setting, because there is high overdispersion, resulting from a relatively small α . This implies that there is a strong source of measurement error, inhibiting a high level of heritability.

When this overdispersion is not taken into account, different estimates for heritability can be found. This is the case of considering the model (W-N) instead of (WGN). To calculate the heritability for this case we apply the expression (2.7) using (2.12). Based on the estimates for the (W-N) we obtain $\zeta_{WN} = 0.9451$ and:

$$H^2 = 1 - 0.9451 = 0.0549(\text{s.e. } 0.0234).$$

Both estimates are similar in view of a relatively large standard error. This is not surprising given the use of the delta method for a ratio. Also, the estimates are close the boundary of

the quantity's range, adding to the complexity to obtain a narrow standard error. Alternatively, Fieller's intervals could be calculated to gauge the estimates' precision.

2.7 Concluding Remarks

In this paper, we have derived a principled expression for heritability, based on hierarchical time-to-event data. To these data, the so-called combined model is fitted, bringing together a generalized linear model for time-to-event data with both normal and gamma random effects. The normal random effects, entering in the linear predictor, account for correlation among dependent measures as well as for some overdispersion, while the gamma random effects, entering in a conjugate way, take care of residual overdispersion. Importantly, the standard generalized linear mixed model is a special case of the combined model, implying that the derivations reported here also apply to this commonly encountered GLMM case.

The combined model and its GLMM sub-model admits closed-form expressions for means, variances, and higher-order moments. As a result, correlation coefficients and variance ratios have explicit expressions too. An appealing feature of this model is that the ensuing heritability coefficient does not depend on fixed effects. While, unsurprisingly, the coefficient is not as simple as when outcomes are Gaussian, it is nevertheless sufficiently simple and appealing, in particular in special cases, to be of practical value. Its calculation poses no difficulty.

All analysis were performed using the SAS NLMIXED procedure due to its flexibility to deal with non-normal responses and because it allows the user to specify the likelihood function to their own taste and needs. This procedure, while flexible, comes with its limitations. For example, incorporating a relationship matrix is less straightforward. Should this be necessary, then alternative implementations have to be considered.

References

- ALBUQUERQUE, L.G.; MEYER, K. Estimates of direct and maternal genetic effects for weights from birth to 600 days of age in nelore cattle. **Journal of Animal Breeding and Genetics**, Berlin, v. 118, n.2, p. 83-92, 2001.
- BRESLOW, N. E. Extra-poisson variation in log-linear models. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, London, v. 33, n.1, p. 38-44, 1984.
- BRESLOW, N. E.; CLAYTON, D. G. Approximate inference in generalized linear mixed models. **Journal of the American Statistical Association**, Boston, v. 88, n.421, p. 9-25, 1993.

CARVALHEIRO, R.; PIMENTEL, E. C. G.; CARDOSO, V.; QUEIROZ, S. A.; FRIES, L. A. Genetic effects on preweaning weight gain of nelore-hereford calves according to different models and estimation methods. **Journal of Animal Science**, Champaign, v. 84, p. 2925-2933, 2006.

DUCHATEAU, L.; JANSSEN, P. **The Frailty Model**. New York: Springer, 2008. 316 p.

FALCONER, D. S.; MACKAY, T. F. C. **Introduction to Quantitative Genetics**. 4th ed. Harlow, England: Addison Wesley Longman, 1996. 464 p.

GILMOUR, A. R.; GOGEL, B. J.; CULLIS, B. R.; THOMPSON, R. **ASReml User Guide Release 3.0**. Hemel Hempstead, HP1 1ES, UK: VSN International Ltd, 2009. Disponível em: Www.vsnl.co.uk. Acesso em: 28 maio 2014.

GIOLO, S. R.; DEMÉTRIO, C. G. B. A frailty modeling approach for parental effects in animal breeding. **Journal of Applied Statistics**, Abingdon, v. 38, n. 3, p. 619-629, 2011.

HINDE, J.; DEMÉTRIO, C. G. B. Overdispersion: Models and estimation. **Computational Statistics & Data Analysis**, Amsterdam, v. 27, n. 2, p. 151-170, 1998.

LAWLESS, J. F. Negative binomial and mixed poisson regression. **The Canadian Journal of Statistics**, Ottawa, v. 15, n. 3, p. 209-225, 1987.

MCCULLAGH, P.; NELDER, J. A. **Generalized linear models**. London: Chapman & Hall, 1989. 532 p.

MOLENBERGHS, G.; VERBEKE, G. **Models for discrete longitudinal data**. New York: Springer, 2005. 683 p.

MOLENBERGHS, G.; VERBEKE, G. On the weibull-gamma frailty model, its infinite moments, and its connection to generalized log-logistic, logistic, cauchy, and extreme-value distributions. **Journal of Statistical Planning and Inference**, Amsterdam, v. 141, n. 2, p. 861-868, 2011.

MOLENBERGHS, G.; VERBEKE, G.; DEMÉTRIO, C. G. B. An extended random-effects approach to modeling repeated, overdispersed count data. **Lifetime Data Analysis**, Hingham, v. 13, n. 4, p. 513-531, 2007.

MOLENBERGHS, G.; VERBEKE, G.; DEMÉTRIO, C. G. B.; VIEIRA, A. A family of generalized linear models for repeated measures with normal and conjugate random effects. **Statistical Science**, Hayward, v. 25, n. 3, p. 325-347, 2010.

NEVES, H. H. R.; CARVALHEIRO, R.; QUEIROZ, S. A. Genetic and environmental heterogeneity of residual variance of weight traits in nellore beef cattle. **Genetics, Selection, Evolution**, Les Ulis, v. 44, n. 19, p. 1-12, 2012.

SCHUKKEN, Y. H.; BAR, D.; HERTL, J.; GRÖHN, Y. T. Correlated time to event data: Modeling repeated clinical mastitis data from dairy cattle in New York state. **Preventive Veterinary Medicine**, Amsterdam, v. 97, n. 3-4, p. 150-156, 2010.

VANGENEUGDEN, T.; MOLENBERGHS, G.; VERBEKE, G.; DEMÉTRIO, C.G.B. Marginal correlation from an extended random-effects model for repeated and overdispersed counts. **Journal of Applied Statistics**, Abingdon, v. 38, n. 2, p. 215-232, 2011.

VAUPEL, J. W.; MANTON, K. G.; STALLARD, E. The impact of heterogeneity in individual frailty on the dynamics of mortality. **Demography**, Chicago, v. 16, n. 3, p. 439-454, 1979.

VERBEKE, G.; MOLENBERGHS, G. **Linear mixed models for longitudinal data**. New York: Springer-Verlag, 2000. 568 p.

VISSCHER, P. M.; HILL, W. G.; WRAY, N. R. Heritability in the genomics era - concepts and misconceptions. **Nature Reviews, Genetics**, London, v. 9, p. 255-266, 2008.

WELSH, A. H. **Aspects of Statistical Inference**. Hoboken: Wiley Series in Probability and Statistics, 2011. 480 p.

3 QUANTIFYING THE GENETIC CONTRIBUTION TO THE VARIABILITY OF COUNT TRAITS

Abstract

Heritability is an important concept in animal and plant breeding, as it is in human biological applications. It is quantified based on fitting a model to hierarchical data. For data where linear models can be used, this attribute is conveniently defined as a ratio of variance components. Matters are less simple for non-Gaussian outcomes. The focus here is on count outcomes where extensions of the Poisson model are used to describe the data. Expressions for heritability of count traits are derived using the so-called Poisson combined model, which combines a Poisson outcome distribution with normal as well as gamma random effects to capture both correlation among repeated observations as well as overdispersion, and admits closed-form expressions for the mean, variances and, hence, ratio of variances. It thus flexibly accommodates overdispersion and within-unit correlation. The proposed methodology is illustrated using maize data from a plant breeding program and compared with the usual, but questionable analysis using linear mixed models.

Keywords: Combined model; Gamma distribution; Generalized linear mixed model; Overdispersion; Poisson distribution; Random effect

3.1 Introduction

Plant breeding is the purposeful manipulation of plant species to improve certain aspects of plants, so as to perform new roles or enhance existing ones. Heritability, defined as the proportion of the genetic contribution over the total variability in a phenotype, is often of importance in plant and animal breeding. Knowledge of this attribute is useful to quantify the magnitude of improvement in the population and it is used when predicting the outcome of selection practiced among clones, inbred lines, or varieties (HARTL; JONES, 2009).

The heritability determination is routinely based on hierarchical data of a family-based nature. When the outcomes are normally distributed, linear mixed models (VERBEKE; MOLENBERGHS, 2000) are frequently used to estimate the genetic and environmental effects by considering these factors as random terms in the model. For data where these models can be used, the heritability can be quantified as the ratio of the genotypic variance, σ_g^2 say, to the total phenotypic variance, σ_p^2 say.

However, when the trait of interest is not normally distributed and/or it does not follow a linear model, the genetic and environmental random terms are no longer easily separable from the other model terms. This difficulty arises in particular when one deals with count outcomes, which are common in agricultural and livestock studies and are the focus of this paper. One often models such data using generalized linear modeling (MCCULLAGH; NELDER, 1989), which covers widely used statistical models, such as Poisson log-linear models for count data.

In empirical research, it has been observed recurrently that the mean-variance relationship

for Poisson model may not be met. As a result, quite a bit of research was devoted to formulate models to deal with this phenomenon, referred to as overdispersion or, sometimes also occurring, underdispersion. We will simply refer to overdispersion.

The so-called generalized linear mixed model (GLMM) (BRESLOW; CLAYTON, 1993; MOLENBERGHS; VERBEKE, 2005) has gained popularity in discrete hierarchical data modeling. When overdispersion and the need for hierarchical modeling occur simultaneously, the combined models proposed by Molenberghs et al. (2010) can be used. These models accommodate overdispersion and clustering through two separate sets of random effects and contain as special cases the GLMM on the one hand, and several overdispersion models, such as the negative-binomial model, on the other. In this paper, we use such models for handling overdispersion and correlated data, while obtaining heritability coefficients based on count traits.

The proposed methodology will be illustrated using data from a study in plant breeding. The data refer to the number of tassel branches in maize progenies. This count trait is of interest in breeding procedures, with smaller numbers considered better. We derive heritability using two approaches: (a) linear mixed models, a conventional but in principle inadequate analysis, and (b) using Poisson models.

The paper is organized as follows. In Section 3.2, the motivating case is described with both analyses reported in Section 3.5. A review of the Poisson combined model for hierarchical and overdispersed count data is the subject of Section 3.3. We use this combined model to obtain heritability coefficients for count traits, the expressions of which are presented in Section 3.4.

3.2 Motivating Case Study

The data considered here are obtained from a square lattice designed experiment, implemented in four different environments, for the selection of maize progenies in a plant breeding program at ESALQ, Piracicaba, Brazil. Forty-nine progenies were replicated twice in each environment, which consists of a combination between a crop season (first *versus* second crop season) and a location (ESALQ *versus* Sertãozinho). The planting during the first crop season usually occurs between September and December, while the planting in the second crop season takes place between January and April. This division of planting periods is common in tropical countries such as Brazil while in temperate regions there is only one planting season.

Five plants were randomly selected from each plot and several phenotypic characteristics were measured, including the kernel-row number per ear, kernel number per row and the number of tassel branches. The latter is a characteristic related to drought tolerance; the smaller the number of branches, the better. This is because the plant does not need to move so many photo-assimilates there rather than move to the ear. On the other hand, the kernel-row number and kernel number per row are components of production, generally with higher values preferred.

The key research questions are what proportion of the variability of each of these traits is

Figure 3.1 – Maize Data. Frequency plots for number of tassel branches, over both replications, over all progenies, and in each environment

due to the genetic effect, i.e., the heritability for this population, and what progenies present the best predicted values with respect to these characteristics. As a summary of the data with respect to the traits aforementioned, Figures 3.1, 3.2 and 3.3 show the frequency plots, over both replications, over all progenies, and in each environment.

In general, the number of branches varies from 7 to 29 and the distribution is approximately symmetric. The kernel-row number is always an even number that varies from 10 to 20 in this population.

3.3 An Extended Poisson Model to Handle Hierarchical and Overdispersed Data

In certain applications of standard generalized linear models, it is found that the data exhibit overdispersion, i.e., the variability is greater than predicted by the mean-variance relationship inherent in the model formulation. A number of models have been proposed for handling this phenomenon, especially in the Poisson case (BRESLOW, 1984; LAWLESS, 1987). Some also handle the rarer case of underdispersion.

An elegant formulation is through a two-stage approach. In the univariate Poisson case, we assume that $Y_i|\lambda_i \sim \text{Pois}(\lambda_i)$ and then that λ_i is a random variable with $E(\lambda_i) = \mu_i$ and

Figure 3.2 – Maize Data. Frequency plots for kernel-row number per ear, over both replications, over all progenies and in each environment

$\text{Var}(\lambda_i) = \sigma_i^2$. Then it follows that

$$E(Y_i) = E[E(Y_i|\lambda_i)] = E(\lambda_i) = \mu_i,$$

$$\text{Var}(Y_i) = E[\text{Var}(Y_i|\lambda_i)] + \text{Var}[E(Y_i|\lambda_i)] = E(\lambda_i) + \text{Var}(\lambda_i) = \mu_i + \sigma_i^2.$$

It is common to assume a gamma distribution for the random effects λ_i , leading to the so-called negative-binomial model.

This model can be extended to the case of repeated measurements. We then assume a hierarchical data structure where Y_{ij} denotes the j th outcome measured for cluster i , ($i = 1, \dots, N; j = 1, \dots, n_i$) and $\mathbf{Y}_i$ is the n_i -dimensional vector of all measurements available for cluster i . The scalar λ_i becomes a vector $\boldsymbol{\lambda}_i = (\lambda_{i1}, \dots, \lambda_{in_i})'$, with $E(\boldsymbol{\lambda}_i) = \boldsymbol{\mu}_i$ and $\text{Var}(\boldsymbol{\lambda}_i) = \Sigma_i$. Then, $E(\mathbf{Y}_i) = \boldsymbol{\mu}_i$ and $\text{Var}(\mathbf{Y}_i) = M_i + \Sigma_i$, where M_i is a diagonal matrix with the vector $\boldsymbol{\mu}_i$ along the main diagonal. The diagonal structure of M_i reflects the conditional independence assumption, that is, all dependence between measurements on the same unit stems from the random effects.

Alternatively, this repeated version of the overdispersion model can be combined with normal random effects in the linear predictor. Such models, proposed also by Thall and Vail (1990) and Dean (1991), for the count case, will be discussed next.

Molenberghs et al. (2007) specified a model for count data combining ideas from the

Figure 3.3 – Maize Data. Frequency plots for kernel number per row, over both replications, over all progenies and in each environment

overdispersion models and models with normal random effects. Later, Molenberghs et al. (2010) proposed a broad class of generalized linear models where the binary, count, and time-to-event cases were given particular emphasis. These models, named combined models, accommodate overdispersion and clustering through two separate sets of random effects and produce models with only random effects and models with only overdispersion as special cases.

Molenberghs et al. (2007) showed that the count models allow for closed-form expressions for the mean vector and variance-covariance matrix. As highlighted by the authors, the derivation of such closed forms has important implications because they admit, for example, explicit correlation expressions. This aspect was examined by Vangeneugden et al. (2011) for Poisson-type models for count data.

In line with Booth et al. (2003), Molenberghs et al. (2007) specified a model for repeated Poisson data with overdispersion:

$$\begin{aligned}
 Y_{ij} &\sim \text{Poi}(\theta_{ij}\lambda_{ij}), \\
 \lambda_{ij} &= \exp(x'_{ij}\boldsymbol{\xi} + z'_{ij}\mathbf{b}_i), \\
 \mathbf{b}_i &\sim N(\mathbf{0}, D), \\
 E(\boldsymbol{\theta}_i) &= E[(\theta_{i1}, \dots, \theta_{in_i})'] = \boldsymbol{\Phi}_i, \\
 \text{Var}(\boldsymbol{\theta}_i) &= \boldsymbol{\Sigma}_i.
 \end{aligned} \tag{3.1}$$

Notice that the normal random effect to capture correlation among repeated observations is placed in the linear predictor while the random effect to accommodate overdispersion acts multiplicatively in the mean of the variable.

The θ_{ij} can be assumed to follow a gamma model producing a Poisson-Gamma-Normal model or, equivalently, a Negative-Binomial-Normal model. When the gamma distribution is chosen, it is assumed that the components θ_{ij} of $\boldsymbol{\theta}_i$ are independent. In this case, Σ_i reduces to a diagonal matrix. It should be noted that it is possible to allow for general covariance structures; this is not considered further here.

Then, regarding the overdispersion random effects, three instances could be of interest: (1) the random-effects θ_{ij} are independent; (2) the θ_{ij} are allowed to be dependent; (3) they are equal to each other and hence reduce to $\theta_{ij} = \theta_i$. Independent θ_{ij} imply that the use is strictly confined to capture additional overdispersion, i.e., not captured by the normal random effects. In contrast, when they are allowed to be correlated, they offer a way to model, for example, serial correlation.

The marginal mean vector and variance-covariance matrix were derived by Molenberghs et al. (2007) and are reproduced in Appendix D. We considered this model to calculate the genetic contribution to the total variability of the traits of interest, that is, the heritability. The derivation of such measure is presented in the following section.

3.4 Derivation of Heritability for Count Data

Consider the Poisson-Gamma-Normal model and its variance presented in (17). Also, without loss of generality, set $E(\boldsymbol{\theta}_i) = 1$. The variance is

$$\text{Var}(\mathbf{Y}_i) = \mu_{ij} + \mu_{ij}(P_{i,jj} - 1)\mu_{ij}, \quad (3.2)$$

where

$$\mu_{ij} = \exp\left(x'_{ij}\boldsymbol{\xi} + \frac{1}{2}z'_{ij}Dz_{ij}\right) = \mu_{0ij}\mu_{1ij},$$

with notation as in (3.1), and

$$P_{i,jj} = \mu_{1ij}(\sigma_{i,jj} + 1)\mu_{1ij} = \mu_{1ij}^2(\sigma_{i,jj} + 1).$$

The non-genetic contribution over the total variability is:

$$\zeta_{ij} = \frac{1 + \mu_{0ij}[(\sigma_{i,jj} + 1) - 1]}{\mu_{1ij}\{1 + \mu_{0ij}\mu_{1ij}[\mu_{1ij}^2(\sigma_{i,jj} + 1) - 1]\}}. \quad (3.3)$$

The ratio in (3.3) places the variance presented in (3.2), for $D = 0$, in the numerator and the full variance of the combined model in the denominator. The heritability, that is, the proportion of the total variability related to the genetic effect is:

$$H_{ij}^2 = 1 - \zeta_{ij}. \quad (3.4)$$

The full variance is the phenotypic variance, σ_p^2 , and the obtained heritability is at individual level. Because of the mean-variance relationship, ζ_{ij} and hence H_{ij}^2 depends on the mean, thus also on the covariates.

A specific case arises when there is no overdispersion. In such case, the non-genetic contribution can be derived from the Poisson-Normal model or simply by setting $\sigma_{i,jj} = 0$ in (3.3). The contribution then is

$$\zeta_{PN_{ij}} = \frac{1}{\mu_{1ij}[1 + \mu_{0ij}\mu_{1ij}(\mu_{1ij}^2 - 1)]}, \quad (3.5)$$

and the heritability is calculated from replacing ζ_{ij} by $\zeta_{PN_{ij}}$ in (3.4).

It is not uncommon to model other sources of variability, such as the environmental effect. This effect can be included as a random term in model (3.1), which slightly changes the terms:

$$\begin{aligned} \lambda_{ij} &= \exp(x'_{ij}\boldsymbol{\xi} + z'_{1ij}\mathbf{v}_i + z'_{2ij}\mathbf{w}_i), \\ \mathbf{v}_i &\sim N(\mathbf{0}, D_1), \\ \mathbf{w}_i &\sim N(\mathbf{0}, D_2), \end{aligned}$$

where $\mathbf{v}_i$ and $\mathbf{w}_i$ are the genetic and environmental effects, respectively. Then, the variance is:

$$\text{Var}(Y_{ij}) = \mu_{ij} + \mu_{ij}(P_{i,jj} - 1)\mu_{ij},$$

where

$$\mu_{ij} = \exp\left(x'_{ij}\boldsymbol{\xi} + \frac{1}{2}z'_{1ij}D_1z_{1ij} + \frac{1}{2}z'_{2ij}D_2z_{2ij}\right) = \mu_{0ij}\mu_{1ij}\mu_{2ij}$$

and

$$P_{i,jj} = \mu_{1ij}\mu_{2ij}(\sigma_{i,jj} + 1)\mu_{1ij}\mu_{2ij} = \mu_{1ij}^2\mu_{2ij}^2(\sigma_{i,jj} + 1).$$

In this case, the contribution from overdispersion and from the random effect $\mathbf{w}_i$ over the total variability is:

$$\zeta_{ij} = \frac{1 + \mu_{0ij}\mu_{2ij}[\mu_{2ij}(\sigma_{i,jj} + 1) - 1]}{\mu_{1ij}\{1 + \mu_{0ij}\mu_{1ij}\mu_{2ij}[\mu_{1ij}^2\mu_{2ij}^2(\sigma_{i,jj} + 1) - 1]\}}. \quad (3.6)$$

When there is no overdispersion, (3.6) reduces to:

$$\zeta_{PN_{ij}} = \frac{1 + \mu_{0ij}\mu_{2ij}(\mu_{2ij} - 1)}{\mu_{1ij}[1 + \mu_{0ij}\mu_{1ij}\mu_{2ij}(\mu_{1ij}^2\mu_{2ij}^2 - 1)]}. \quad (3.7)$$

The heritability values are obtained from applying (3.6) or (3.7) to (3.4).

Notice that the ratios ζ_{ij} and $\zeta_{PN_{ij}}$ are not free of the marginal mean function. In practice, therefore, one should compute some useful summaries of the values H_{ij} , given that they depend on the means at measurement j for unit i . Of course, when covariates are limited to a few factors with a limited number of levels, like is the case in our applications, the heritability coefficients will only be dependent on these.

3.5 Data Analysis

The data introduced in Section 3.2 were analyzed using two approaches. First, we used linear models, which are frequently used for genetic evaluation, referred to as ‘conventional analysis.’ In spite of its convenience, we consider it less adequate because it does not do full justice to the type of data collected. Second, we used the appropriate methodology from the previous section. From the estimates obtained under both approaches, we calculated the heritability for the number of tassel branches in this population.

It is noteworthy that, in this work, we do not model the progeny-environment interaction that can exist in this kind of study. So, all analyses were performed separately for each environment. Furthermore, we opted to work with the observations only at the plot level, which is the lowest and most informative hierarchical level, and did not estimate the block and replication nested effects of the lattice design.

The following linear predictor was considered in both approaches:

$$\eta_{ij} = \xi_0 + b_i,$$

where ξ_0 is an effect common to all observations, b_i is the genetic random effect of the i th progeny (49 progenies), and $b_i \sim N(0, \sigma_g^2)$. In the linear models, a random residual effect, ε_{ij} say, completes the model specification, where $\varepsilon_{ij} \sim N(0, \sigma^2)$.

3.5.1 The Conventional Linear Analysis

In this section, we will analyze the number of tassel branches considering the linear fixed effects and linear mixed effects models. Results from fitting both models are displayed in Table 3.1. Unsurprisingly, the goodness-of-fit of the mixed models is higher than that of univariate linear regression. From the estimated variance components and using the classical definition of heritability at the individual level, $H^2 = \sigma_g^2 / (\sigma_r^2 + \sigma_g^2)$, we calculated the genetic contribution in each one of the environments.

The heritability values are $H^2 = 0.1931$ for the first crop season at ESALQ, $H^2 = 0.3864$ for the first crop season at Sertãozinho, $H^2 = 0.3508$ for the second crop season at ESALQ, and $H^2 = 0.266$ for the second crop season at Sertãozinho.

3.5.2 The Proposed Poisson-type Analysis

In this section, we will analyze the number of tassel branches considering (1) the Poisson model (P--), (2) the Poisson Normal (P-N), (3) the Poisson-Gamma model (PG-), (4) the Poisson-Gamma-Normal model (PGN). Results from fitting these models are displayed in Table 3.2.

The estimates of the Gamma parameter were very high in all environments, demonstrating the absence of overdispersion in this application. Again, as expected, in all cases, the goodness-

Table 3.1 – Maize study. Parameter estimates and standard errors for the regression coefficients in (1) the linear model (LM) and (2) linear mixed model (LMM), considering the trait number of tassel branches in the four environments

Effect	Par.	(LM)	(LMM)
1. First crop season at ESALQ			
Intercept	ξ_0	16.359 (0.146)	16.359 (0.241)
Var. of progenies	σ_g^2		2.013 (0.579)
Var. residual	σ_r^2	10.410 (0.665)	8.397 (0.566)
-2log-likelihood		2538.5	2493.2
AIC		2542.5	2499.2
2. First crop season at Sertãozinho			
Intercept	ξ_0	16.571 (0.162)	16.571 (0.344)
Var. of progenies	σ_g^2		4.995 (1.171)
Var. residual	σ_r^2	12.931 (0.826)	7.935 (0.534)
-2log-likelihood		2644.8	2502.9
AIC		2648.8	2508.9
3. Second crop season at ESALQ			
Intercept	ξ_0	15.157 (0.138)	15.157 (0.282)
Var. of progenies	σ_g^2		3.292 (0.789)
Var. residual	σ_r^2	9.381 (0.599)	6.090 (0.410)
-2log-likelihood		2487.5	2366.8
AIC		2491.5	2372.8
4. Second crop season at Sertãozinho			
Intercept	ξ_0	15.818 (0.154)	15.818 (0.284)
Var. of progenies	σ_g^2		3.105 (0.803)
Var. residual	σ_r^2	11.687 (0.747)	8.582 (0.578)
-2log-likelihood		2595.2	2518.9
AIC		2599.2	2524.9

of-fit of the (P-N) models is higher than that of the (P--) models. So, the (P-N) estimates were used in (3.5) and (3.4) to obtain heritability.

The estimated non-genetic contributions and heritability values are $\zeta_{PN} = 0.9219$ and $H^2 = 0.0781$ for the first crop season at ESALQ, $\zeta_{PN} = 0.7828$ and $H^2 = 0.2172$ for the first crop season at Sertãozinho, $\zeta_{PN} = 0.8635$ and $H^2 = 0.1365$ for the second crop season at ESALQ and $\zeta_{PN} = 0.8586$, and $H^2 = 0.1414$ for the second crop season at Sertãozinho.

Another question of interest in plant breeding programs is to identify the best progeny with respect to the characteristic under investigation, and, for this, we plotted the empirical Bayes estimates for the random effects (Figure 3.4). In all environments, progeny 45 has the lowest predicted number of tassel branches, except for the second crop season at ESALQ, where progeny 49 showed superior performance.

Note that the heritability values calculated from the (P-N) models are smaller than those obtained from the linear-mixed model analysis. Close attention should be given to this point: a misspecification of the data distribution can lead to overestimated random effects and, hence, result in erroneously high heritability values.

It is clear that the original structure of the lattice design was not considered and a completely

randomized design model was used. For purposes of illustration, it was an acceptable choice. By doing so, the blocks and replications effects were not modeled and their variation was added to the variance component σ_r^2 , which is overestimated. As a result, the estimated heritability is underestimated. Even the heritabilities for the linear mixed model are underestimated, they are higher than those considering the Poisson distribution.

To illustrate the effect of model misspecification in the estimation of genetic quantities, we also analyzed the kernel-row number per ear. The estimates of the models considered are presented in Appendix F. For this trait, there is a significant effect of the ear diameter covariate, considering the linear mixed models in all environments and the Poisson model in the first crop season at Sertãozinho. Also for this environment, the (PG-) and (PGN) models failed to converge, but this is not a surprise due to the high estimates of the Gamma parameter and null estimates of random effects.

If one assumes the normal distribution for this count trait, the heritability values are $H^2 = 0.0806$ for the first crop season at ESALQ, $H^2 = 0.1640$ for the first crop season at Sertãozinho, $H^2 = 0.2936$ for the second crop season at ESALQ and $H^2 = 0.1590$ for the second crop season at Sertãozinho. On the other hand, considering the Poisson models, there are neither significant overdispersion nor significant normal random effects and, hence, the heritability is not significantly different from 0.

The amount of genetic variation determines the rate of change of a trait under selection and if there is no genetic variation, there is no response to selection (HARTL; CLARK, 2007).

Table 3.2 – Maize study. Parameter estimates and standard errors for the regression coefficients in (1) the purely Poisson model (P--), (2) the Poisson-Normal model (P-N), (3) the Poisson-Gamma model (PG-), and (4) the Poisson-Gamma-Normal model (PGN), considering the trait number of tassel branches in the four environments

Effect	Par.	(P--)	(P-N)	(PG-)	(PGN)
1. First crop season at ESALQ					
Intercept	ξ_0	2.795 (0.011)	2.793 (0.015)	2.795 (0.011)	2.793 (0.015)
Gamma param.	α			1000.020 (580.310)	7638.150 (11268)
Var. of progenies	σ_g^2		0.005(0.002)		0.005 (0.002)
-2log-likelihood		2573.6	2564.4	-28771	-28783
AIC		2575.6	2568.4	-28767	-28777
2. First crop season at Sertãozinho					
Intercept	ξ_0	2.808 (0.011)	2.800 (0.021)	2.808 (0.011)	2.800 (0.021)
Gamma param.	α			5982.350 (10830)	7735.860 (11025)
Var. of progenies	σ_g^2		0.016 (0.005)		0.016 (0.005)
-2log-likelihood		2658.3	2595.4	-29.356	-29419
AIC		2660.3	2599.4	-29352	-29413
3. Second crop season at ESALQ					
Intercept	ξ_0	2.719 (0.012)	2.713 (0.019)	2.719 (0.012)	2.713 (0.019)
Gamma param.	α			6786.550 (10475)	7791.040 (10969)
Var. of progenies	σ_g^2		0.010 (0.004)		0.010 (0.004)
-2log-likelihood		2532.4	2501.6	-25526	-25556
AIC		2534.4	2505.6	-25522	-25550
4. Second crop season at Sertãozinho					
Intercept	ξ_0	2.761 (0.011)	2.756 (0.018)	2.761 (0.011)	2.756 (0.018)
Gamma param.	α			6008.200 (10200)	7660.940 (12063)
Var. of progenies	σ_g^2		0.010 (0.003)		0.010 (0.003)
-2log-likelihood		2614.3	2584.9	-27301	-27331
AIC		2616.3	2588.9	-27297	-27325

Figure 3.4 – Maize Data. Empirical Bayes Estimates for number of tassel branches in each environment

3.6 Concluding Remarks

In this paper, we have derived a principled expression for heritability, based on hierarchical count data. To these data, Poisson-based mixed models as well as linear mixed models have been fitted and compared. Although conventional, we consider inadequate the use of linear (mixed) models for count data and showed that, using this approach, the genetic random effects can be overestimated, leading to incorrect heritability values. Of course, when counts would be very large and at least approximately normally distributed, the above judgment can be relaxed.

The so-called combined model was used in the Poisson approach. It brings together a generalized linear model for count data with both normal and gamma random effects, thus accommodating correlation and overdispersion. Importantly, the standard generalized linear mixed model is a special case of the combined model, implying that the derivations reported here also apply to this commonly encountered GLMM case.

The combined model and its GLMM sub-model admit closed-form expressions for means, variances, and higher-order moments. As a result, variance ratios have explicit expressions too. The heritability coefficient is sufficiently simple and appealing, in particular in special cases, to be of practical value.

We want to reiterate that, in these models, heritability is a function rather than a constant. At first sight, this is a drawback. However, it is a consequence from the mean-variance relationship

in the models considered. If the model fits the data well, it can also be claimed to be a feature of the data. Practically, heritability changes with the effects present in the predictor functions. Evidently, one can summarize the functions in a variety of ways, using averages, medians, quartiles, ranges, etc.

References

BOOTH, J. G.; CASELLA, G.; FRIEDL, H.; HOBERT, J. P. Negative binomial loglinear mixed models. **Statistical Modelling**, London, v. 3, n. 3, p. 179-191, 2003

BRESLOW, N. E. Extra-poisson variation in log-linear models. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, London, v. 33, n. 1, p. 38-44, 1984.

BRESLOW, N. E.; CLAYTON, D. G. Approximate inference in generalized linear mixed models. **Journal of the American Statistical Association**, Boston, v. 88, n. 421, p. 9-25, 1993.

DEAN, C. B. Estimating equations for mixed-poisson models. In: GODAMBE, V. (Ed.), **Estimating Functions**. Oxford: Oxford University Press, 1991. p. 35-46.

HARTL, D. L.; CLARK, A. G. **Principles of Population Genetics**. 4th ed. Sunderland: Sinauer Associates Associates, 2007. 652 p.

HARTL, D. L.; JONES, E. W. **Genetics: analysis of genes and genomes**. 7th ed. Sudbury, Mass : Jones and Bartlett Pub, 2009. 763 p.

LAWLESS, J. F. Negative binomial and mixed poisson regression. **The Canadian Journal of Statistics**, Ottawa, v. 15, n. 3, p. 209-225, 1987.

MCCULLAGH, P.; NELDER, J. A. **Generalized linear models**. London: Chapman & Hall, 1989. 532 p.

MOLENBERGHS, G.; VERBEKE, G. **Models for discrete longitudinal data**. New York: Springer, 2005. 683 p.

MOLENBERGHS, G.; VERBEKE, G.; DEMÉTRIO, C. G. B. An extended random-effects approach to modeling repeated, overdispersed count data. **Lifetime Data Analysis**, Hingham, v. 13, n. 4, p. 513-531, 2007.

MOLENBERGHS, G.; VERBEKE, G.; DEMÉTRIO, C. G. B.; VIEIRA, A. A family of generalized linear models for repeated measures with normal and conjugate random effects. **Statistical Science**, Hayward, v. 25, n. 3, p. 325-347, 2010.

THALL, P. F.; VAIL, S. C. Some covariance models for longitudinal count data with overdispersion. **Biometrics**, Washington, v. 46, n. 3, p. 657-671, 1990.

VANGENEUGDEN, T.; MOLENBERGHS, G.; VERBEKE, G.; DEMÉTRIO, C. G. B. Marginal correlation from an extended random-effects model for repeated and overdispersed counts. **Journal of Applied Statistics**, Abingdon, v. 38, n. 2, p. 215-232, 2011.

VERBEKE, G.; MOLENBERGHS, G. **Linear mixed models for longitudinal data**. New York: Springer-Verlag, 2000. 568 p.

4 NEGATIVE VARIANCE COMPONENTS FOR NON-NEGATIVE HIERARCHICAL DATA

Abstract

The concept of negative variance components in linear mixed-effects models, while confusing at first sight, has received considerable attention in the literature, for well over half a century, following the early work of Nelder (1954) and Chernoff (1954). Broadly, negative variance components in linear mixed models are allowable if inferences are restricted to the implied marginal model. When a hierarchical view-point is adopted, the variance-covariance matrix of the random effects must be positive (semi-)definite. Many contemporary software packages allow for this distinction. Less work has been done for generalized linear mixed models. Here, we study such models, with extension to allow for overdispersion, for non-negative outcomes (counts). Using a plant breeding study, it is illustrated how such negative variance components play a natural role in modeling both the correlation between repeated measures on the same experimental unit and overdispersion. In particular, when overdispersion is present but modest, and intra-unit correlation is large, negative variance components are needed to properly accommodate this situation.

Keywords: Gamma distribution; Generalized linear mixed model; Overdispersion; Poisson distribution

4.1 Introduction

The need for inference on variance components arises in a variety of applied fields. Existing tools to this effect encompass random-effects ANOVA models (NELDER, 1954), linear mixed models (VERBEKE; MOLENBERGHS, 2000), generalized linear mixed models (MOLENBERGHS; VERBEKE, 2005), and generally models that accommodate overdispersion and clustering (BRITTON, 1997). By definition, when variance components are interpreted as variances, they are non-negative quantities, but the occurrence of negative estimates is a reasonably well understood phenomenon in the context of linear models for hierarchical data. Of course, then the interpretation as variance is dropped; rather, such components play a role in the induced marginal model only.

In studies involving grouped data, it is common that the observations within the same cluster are dependent, which implies that such observations belonging to the same cluster are more similar to one another than observations from different clusters. Such dependence is sometimes measured by the intraclass correlation. Occasionally, though, observations within clusters may be dissimilar, e.g., when there are competition effects. An example of negative influence in grouped data is when there is a fixed source and cluster members have to compete for this resource, leading to a negative intraclass correlation (KENNY et al., 2002), such as when plants in the same plot compete for the same nutrients resources, water and/or light. The resulting negative correlation is then captured by negative variance-component estimates.

The occurrence of negative variance components in linear mixed models was reviewed by Molenberghs and Verbeke (2011). As alluded to before, whenever inference for variance components is required, one will have to make a choice between a hierarchical and a marginal view. Under a marginal interpretation, the variance component can be negative as long as the resulting marginal variance-covariance matrix of the observations be positive definite. On the other hand, when a hierarchical view is adopted, random effects retain their interpretation and, hence, their variances must be non-negative. Pryseley et al. (2011) reviewed the linear case, but the focus was on negative variance components for non-Gaussian data, specifically on binary and count outcomes.

In this paper, our focus is on variance components of the so-called Poisson combined model, presented by Molenberghs et al. (2007) for modeling overdispersion and cluster-induced correlation in count data through two separate sets of random effects. Assuming gamma and normal distributions for the random effects leads to the Poisson-Gamma-Normal (PGN) model.

The marginal variance is made up of contributions from both random effects, as well as from the mean-variance relationship of the underlying generalized linear model. The counterpart for time-to-event data is the Weibull-Gamma-Normal (WGN) model. However, we will focus on the (PGN), for brevity.

In Section 4.2, we review the count data case, from the simple, purely Poisson, to the (PGN). Important expressions related to these models are presented in Appendix D. The variance components related to the gamma and normal random effects in the (PGN) are further studied in Section 4.3. Comments regarding estimation are provided in Section 4.4. Section 4.5 reports on the application of the (PGN) model, from which negative variance components arise.

4.2 Background on Poisson Models

The Poisson model is a natural choice with count data. Such a model is one of the prominent members of the exponential family (MCCULLAGH; NELDER, 1989). The latter provides an elegant and encompassing mathematical framework within the generalized linear modeling context (NELDER; WEDDERBURN, 1972; MCCULLAGH; NELDER, 1989).

Let Y_i be Poisson distributed with mean λ_i , denoted by $Y_i \sim \text{Poi}(\lambda_i)$. The density can be written as

$$f(y_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} = \exp\{y_i \ln \lambda_i - \lambda_i - \ln y_i!\},$$

The variance function equals $\nu_i(\mu_i) = \mu_i = \lambda_i$. The logarithm is the natural link function, leading to the classical Poisson regression model with $\ln \lambda_i = \mathbf{x}_i \boldsymbol{\xi}$.

The model imposes equality of mean and variance, although empirical research has abundantly demonstrated that this assumption is not always met. Therefore, a number of extensions have been proposed (BRESLOW, 1984; LAWLESS, 1987). An elegant way to deal with overdispersion, that is, the variability is greater than predicted by the mean-variance

relationship, is through a random-effects approach: $Y_i|\lambda_i \sim \text{Poi}(\lambda_i)$ and then further that λ_i is a random variable with $E(\lambda_i) = \mu_i$ and $\text{Var}(\lambda_i) = \sigma_i^2$. Using iterated expectations, it follows that

$$\begin{aligned} E(Y_i) &= E[E(Y_i|\lambda_i)] = E(\lambda_i) = \mu_i, \\ \text{Var}(Y_i) &= E[\text{Var}(Y_i|\lambda_i)] + \text{Var}[E(Y_i|\lambda_i)] = E(\lambda_i) + \text{Var}(\lambda_i) = \mu_i + \sigma_i^2. \end{aligned}$$

It is common to assume a gamma distribution for λ_i , leading to the Poisson-Gamma (PG-) model, also known as the negative-binomial model.

This model can easily be extended to the case of repeated measures. For this, let us assume a hierarchical data structure, where Y_{ij} denotes the j th outcome measured for cluster i ($i = 1, \dots, N; j = 1, \dots, n_i$) and $\mathbf{Y}_i$ is the n_i -dimensional vector of all measurements available for cluster i . The vector of parameters is then $\boldsymbol{\lambda}_i = (\lambda_{i1}, \dots, \lambda_{in_i})'$, with $E(\boldsymbol{\lambda}_i) = \boldsymbol{\mu}_i$ and $\text{Var}(\boldsymbol{\lambda}_i) = \boldsymbol{\Sigma}_i$. Then, $E(\mathbf{Y}_i) = \boldsymbol{\mu}_i$ and $\text{Var}(\mathbf{Y}_i) = \mathbf{M}_i + \boldsymbol{\Sigma}_i$ where $\mathbf{M}_i$ is a diagonal matrix with the vector $\boldsymbol{\mu}_i$ along the main diagonal. Assuming the components of $\boldsymbol{\lambda}_i$ to be independent, a pure overdispersion model results, without correlation between the repeated measures. Also, assuming $\lambda_{ij} = \lambda_i$, then $\text{Var}(\mathbf{Y}_i) = \mathbf{M}_i + \sigma_i^2 \mathbf{J}_{n_i}$, where $\mathbf{J}_{n_i}$ is an $n_i \times n_i$ dimensional matrix of ones. Such a structure can be seen as a count-data version of compound symmetry.

In hierarchical data modeling, the generalized linear mixed model (GLMM) (BRESLOW; CLAYTON, 1993; MOLENBERGHS; VERBEKE, 2005) has gained popularity in the context of non-Gaussian measures. For the specific case of count data, the parameters become $\lambda_{ij} = \exp(x'_{ij}\boldsymbol{\xi} + z'_{ij}\mathbf{b}_i)$, with $\mathbf{b}_i \sim N(\mathbf{0}, D)$. Owing to the use of the logarithmic link and the normality of the random effects, the mean vector and variance-covariance matrix of $\mathbf{Y}_i$ can be derived in closed form (MOLENBERGHS et al., 2007). The expressions are presented at the end of Appendix D.

An extended version of the aforementioned models was presented by Molenberghs et al. (2007), in line with Booth et al. (2003). These extensions accommodate correlated count data with overdispersion, simultaneously, combining two separate sets of random effects. Assuming normal and gamma distributions for these random effects, the so-called Poisson-Gamma-Normal (PGN) model follows. It yields the Poisson-Normal (P-N), the Poisson-Gamma (PG-), and the purely Poisson model (P--) as special cases. The (PGN) model elements are:

$$\begin{aligned} Y_{ij} &\sim \text{Poi}(\lambda_{ij}), \\ \lambda_{ij} &= \theta_{ij} \exp(x'_{ij}\boldsymbol{\xi} + z'_{ij}\mathbf{b}_i), \\ \mathbf{b}_i &\sim N(\mathbf{0}, D), \\ \theta_{ij} &\sim \text{Gamma}(\alpha, \beta), \\ E(\boldsymbol{\theta}_i) &= E[(\theta_{i1}, \dots, \theta_{in_i})'] = \boldsymbol{\Phi}_i, \\ \text{Var}(\boldsymbol{\theta}_i) &= \boldsymbol{\Sigma}_i. \end{aligned} \tag{4.1}$$

It is implicitly assumed that the components θ_{ij} of $\boldsymbol{\theta}_i$ are independent, which is often natural because the $\mathbf{b}_i$ components induce association between repeated measures, while θ_{ij} capture

additional dispersion. In this case, Σ_i reduces to a diagonal matrix with the variances of the gamma random effects along the main diagonal. These and the components of D play a crucial role in what follows. The parameters α and β in the gamma distribution are not jointly identified. It is therefore customary to impose restrictions, such as constraining the mean or variance.

The (PGN) and its sub-models admit closed-form expressions for means, variances, and higher-order moments. As a result, the correlations too have closed-form expressions (VANGENEUGDEN et al., 2011). All of these are presented in Appendix D.

4.3 The Case of Random Intercepts and Independent Gamma Variables

We focus on the variance components of the (PGN), for the special and important case where the random-effects structure is reduced to random intercepts only, and with a constant mean function, thereby reducing the linear predictor to merely ξ_0 . We will also assume that the gamma variables are identically and independently distributed, that is, Σ_i is a diagonal matrix, with elements α along the diagonal. In such case, the components of the mean vector, $\mu_i = E(\mathbf{Y}_i)$, presented in Appendix D reduce to

$$\mu = \exp\left(\xi_0 + \frac{1}{2}d\right) = e^{\xi_0} e^{\frac{1}{2}d} \equiv \mu_0 e^{\frac{1}{2}d},$$

where d is the scalar version of D in case there is only one normal random effect. To simplify notation, let $\Delta \equiv e^{\frac{1}{2}d}$. Now, we can rewrite the mean, variance, covariance, and correlation expressions of the (PGN), presented in Appendix D, in terms of Δ :

$$\begin{aligned}\mu &= \mu_0 \Delta, \\ \sigma^2 &= \mu_0 \Delta + \mu_0^2 \Delta^2 (\Delta^2 \alpha + \Delta^2 - 1), \\ \rho &= \frac{\mu_0 \Delta (\Delta^2 - 1)}{1 + \mu_0 \Delta (\Delta^2 \alpha + \Delta^2 - 1)},\end{aligned}\tag{4.2}$$

where μ_0 , Δ and α are unknowns. Assume that $\tilde{\mu}$, $\tilde{\sigma}^2$ and $\tilde{\rho}$, the mean, variance and correlation, are given, and solve μ_0 , Δ and α , where:

$$\tilde{\mu} = \mu_0 \Delta, \tag{4.3}$$

$$\tilde{\sigma}^2 = \mu_0 \Delta + \mu_0^2 \Delta^2 (\Delta^2 \alpha + \Delta^2 - 1), \tag{4.4}$$

$$\tilde{\rho} = \frac{\mu_0 \Delta (\Delta^2 - 1)}{1 + \mu_0 \Delta (\Delta^2 \alpha + \Delta^2 - 1)}. \tag{4.5}$$

Some evident conditions are $\tilde{\sigma}^2 \geq 0$ and $1 \geq \tilde{\rho} \geq -\frac{1}{n-1}$, the latter to ensure the matrix be positive definite. Also, write $\tilde{\mu}\tilde{\theta} = \tilde{\sigma}^2$, where $\tilde{\theta}$ is the overdispersion effect. The solution to the

system of equations (4.3)–(4.5) is:

$$\mu_0 = \frac{\tilde{\mu}^2}{\sqrt{\tilde{\mu}^2 + \tilde{\rho}\tilde{\sigma}^2}} = \sqrt{\frac{\tilde{\mu}^3}{\tilde{\mu} + \tilde{\rho}\tilde{\theta}}}, \quad (4.6)$$

$$\alpha = \frac{\tilde{\sigma}^2 - \tilde{\mu} - \tilde{\rho}\tilde{\sigma}^2}{\tilde{\mu}^2 + \tilde{\rho}\tilde{\sigma}^2} = \frac{\tilde{\theta} - (1 + \tilde{\rho}\tilde{\theta})}{\tilde{\mu} + \tilde{\rho}\tilde{\theta}}, \quad (4.7)$$

$$\Delta^2 = 1 + \frac{\tilde{\rho}\tilde{\sigma}^2}{\tilde{\mu}^2} = \frac{\tilde{\mu} + \tilde{\rho}\tilde{\theta}}{\tilde{\mu}}, \quad (4.8)$$

leading to

$$d = \ln\left(\frac{1 + \tilde{\rho}\tilde{\sigma}^2}{\tilde{\mu}^2}\right) = \ln\left(\frac{\tilde{\mu} + \tilde{\rho}\tilde{\theta}}{\tilde{\mu}}\right). \quad (4.9)$$

In the following we will study this solution in some detail.

4.3.1 Variance Component Induced by the Gamma Random Effect

First, we study the variance component related to the gamma random effect, α . If $\tilde{\rho} = 0$, that is, there is no intraclass correlation, $d = 0$, because $\Delta^2 = e^d = 1$, which is obvious given the normal random effect captures the association between repeated measurements.

Turning attention to the gamma variance component where $\tilde{\rho} = 0$, we have that $\alpha = (\tilde{\sigma}^2 - \tilde{\mu})/\tilde{\mu}^2 = (\tilde{\theta} - 1)/\tilde{\mu}$. Hence, if there is no overdispersion, that is, $\tilde{\theta} = 1$, then $\alpha = 0$. If there is overdispersion ($\tilde{\sigma}^2 > \tilde{\mu}$), $\tilde{\theta} > 1$ and $\alpha > 0$. On the other hand, $\alpha < 0$ when $\tilde{\theta} < 1$, that is, when underdispersion occurs ($\tilde{\sigma}^2 < \tilde{\mu}$).

When there is perfect positive intraclass correlation, $\tilde{\rho} = 1$. In such a case, d is positive because $\Delta^2 = 1 + \tilde{\sigma}^2/\tilde{\mu}^2 = 1 + \tilde{\theta}/\tilde{\mu}$. On the other hand, $\alpha = -\tilde{\mu}/(\tilde{\mu}^2 + \tilde{\sigma}^2) = -(\tilde{\mu} + \tilde{\theta})^{-1}$ and its sign depends on whether $\tilde{\mu} + \tilde{\theta} > 0$ or $\tilde{\mu} + \tilde{\theta} < 0$. The latter is not possible, implying that for perfect positive correlation, α must be negative. This underscores that negative values for α are plausible from a practical standpoint. Of course, for real data, the intraclass correlation may be very large, where $\tilde{\rho} = 1$ should be viewed as a limiting case.

Another useful scenario is when there is no overdispersion ($\tilde{\theta} = 1$) and $\tilde{\rho}$ is arbitrary. In such case $\Delta^2 = (\tilde{\mu} + \tilde{\rho})/\tilde{\mu}$, implying $d = \ln[(\tilde{\mu} + \tilde{\rho})/\tilde{\mu}]$ and $\alpha = -\tilde{\rho}(\tilde{\mu} + \tilde{\rho})$. Observe that, if there is a positive intraclass correlation ($\tilde{\rho} > 0$) then α must be negative, implying that $\tilde{\rho} \geq 0$, where $\tilde{\rho} \in [0, 1]$, or $\tilde{\rho} < -\tilde{\mu}$, where $\tilde{\rho} \in [-1, -\tilde{\mu}[$, which can happen only for $\tilde{\mu}$ in the unit interval. Thus, if there is positive correlation and no overdispersion, and because $\tilde{\rho}$ induces extra variance, this must be removed again, leading to negative α .

On the other hand, without overdispersion and for $\tilde{\rho} = 0$, then $\Delta^2 = 1$, $d = 0$ and $\alpha = 0$.

Observing the solution presented in (4.6)–(4.9), it is clear that α can be infinity if $\tilde{\mu} + \tilde{\rho}\tilde{\theta} = 0$, implying that $\tilde{\rho}\tilde{\theta} = -\tilde{\mu}$. Then, μ_0 tends to $+\infty$ and $\Delta^2 = 0$, that is, d tends to $-\infty$.

4.3.2 Variance Component Induced by the Normal Random Effect

The variance component d , associated with the normal random effects in the (PGN) is non-zero if and only if $\Delta^2 \geq 0$. Then,

$$1 + \frac{\tilde{\rho}\tilde{\sigma}^2}{\tilde{\mu}^2} \geq 0 \iff \tilde{\rho} \geq -\frac{\tilde{\mu}^2}{\tilde{\sigma}^2} = -\frac{\tilde{\mu}}{\tilde{\theta}}.$$

Depending on the values of $\tilde{\mu}$ and $\tilde{\theta}$, this will or will not be a real condition. Precisely, if $\tilde{\mu} \geq \tilde{\theta}$ the above condition is sufficient. On the other hand, there is an additional restriction if $\tilde{\mu} \leq \tilde{\theta}$. For d to be non-negative:

$$\begin{aligned} 1 + \frac{\tilde{\rho}\tilde{\sigma}^2}{\tilde{\mu}^2} \geq 1 &\iff \frac{\tilde{\rho}\tilde{\sigma}^2}{\tilde{\mu}^2} \geq 0 \\ &\iff \tilde{\rho} \geq 0, \end{aligned}$$

because $\tilde{\sigma}^2, \tilde{\mu}^2 \geq 0$. So, non-negative intraclass correlation ($\tilde{\rho}$) implies non-negative d and negative intraclass correlation implies negative d .

4.4 Estimation

Standard generalized linear mixed models can be fitted to data with a variety of software tools, such as the SAS procedures GLIMMIX and NLMIXED. These and other tools offer a variety of numerical optimization algorithms, a key component of which is the method for integrating over the normal random effects. It is well known (MOLENBERGHS; VERBEKE, 2005) that Taylor-series-expansion based methods, such as MQL and PQL, perform poorly, especially with binary data, but that the quality of the approximation used, especially for PQL, improves with count and time-to-event data. Further methods are based on Laplace approximations and Gauss-Hermite quadrature. The combined model can generally be fitted using the SAS procedure NLMIXED, because it allows to flexibly use program statements for the conditional likelihood. The conditional likelihood here is understood as the likelihood integrated over the conjugate but not over the normal random effects. Using the example in the next section, we will assess the ease with which boundary and/or negative estimates can be reached using the various methods. When a negative estimate is allowed for by the user, it is still possible that it cannot be found purely because an algorithm is used that does not allow for it. For example, we will note that the Laplace approximation allows for negative normal random-effects variances. There is then a tradeoff between the accuracy of a method on the one hand and its capability of extending the parameter space of the variance components on the other.

4.5 A Maize Breeding Study

We consider a maize breeding study, where the objective is to improve certain aspects of the plants by means of crossing the best genotypes in view of such aspects. The data were obtained

from a square lattice designed experiment, implemented in four different environments, for the selection of maize genotypes in a plant breeding program organized by ESALQ, Piracicaba, Brazil. Forty-nine genotypes were replicated twice in each environment, which consists of a combination between a planting season (first *versus* second crop season) and a location (ESALQ *versus* Sertãozinho).

Five plants were randomly selected from each plot, which consisted of 20 plants of a genotype. Several phenotypic characteristics were measured, including the count traits number of tassel branches (NTB), kernel row number (KR) and kernel number per row (KN).

Some points need to be highlighted for this type of study. First, there is a natural competition among plants within the same plot and between those of adjacent plots. This intra-specific competition occurs for soil resources (water and nutrients) and solar radiation, and can lead to a reduction in grain yield. Secondly, because some crops have been through years of breeding, the genetic differences between individuals are more subtle and, hence, the gains of breeding are less expressive. At some point, it is expected that the overall variability between individuals with respect to a given characteristic is low and, therefore, it is not surprising that there is underdispersion or no overdispersion in some cases. In view of these considerations, negative estimates of variance components can occur in this kind of study.

As noted in Section 4.4, we fitted the (PGN) to the maize data, employing PQL, Laplace, and adaptive Gauss-Hermite quadrature, as implemented in the SAS procedure GLIMMIX. All analyses were performed separately for each environment. Furthermore, we opted for working with the observations only at the plot level, which is the lowest and most informative hierarchical level, and did not estimate the block and replication nested effects of the lattice design.

The results for the PQL method, which showed the best performance for this application in terms of numerical stability, are displayed in Table 4.1. Results from the other estimation methods considered are presented and briefly discussed in Appendix H. Recall that the PQL method can be based on a relatively poor Taylor series approximation, which is a factor to be taken into consideration next to the numerical stability.

The estimates for intraclass correlations displayed in Table 4.1 were obtained from (4.2) and their standard errors were calculated using the delta method (WELSH, 1996). Details on the calculation are presented in Appendix G. Because the correlation expression depends on covariate effects (whenever present), we presented an informal interval for the correlations on all levels of the covariate ear length, for the number of grains per row.

Despite the expected intra-specific competition, in neither case negative correlations were obtained. This can be explained by the fact that the ten specimen of the same genotype (cluster) were grown in two different replications of the experiment. Then, the plots of both replications may be different in one way or another, and their individuals may not influence each other. Furthermore, to evaluate these characters, five plants were randomly selected in the total of twenty plants and such selected plants may not be so close spatially.

Table 4.1 – Maize study. Parameter estimates (standard errors) for the (PGN), considering random intercepts, for number of tassel branches (NTB), kernel row number (KR) and kernel number per row (KN). The estimation method used was penalized quasi-likelihood (PQL)

Effect	Par.	NTB	KR	KN
1. First crop season at ESALQ				
Intercept	ξ_0	2.79 (0.02)	2.62 (6.53×10^{-3})	3.36 (0.06)
Ear diameter	ξ_1			
Ear length	ξ_2			0.02 (3.26×10^{-3})
Overdispersion	α	-0.03 (2.08×10^{-3})	-0.06 (2.79×10^{-4})	-0.02 (5.76×10^{-4})
Compound symmetry	d	7.65×10^{-3} (2.26×10^{-3})	1.10×10^{-3} (3.30×10^{-4})	1.07×10^{-3} (4.14×10^{-4})
Correlation	ρ	0.19 (0.05)	0.10 (0.03)	[0.09 (0.03); 0.13 (0.05)]
2. First crop season at Sertãozinho				
Intercept	ξ_0	2.80 (0.02)		3.31 (0.05)
Ear diameter	ξ_1			
Ear length	ξ_2			0.02 (2.98×10^{-3})
Overdispersion	α	-0.03 (1.87×10^{-3})		-0.02 (6.01×10^{-4})
Compound symmetry	d	0.02 (4.80×10^{-3})		1.06×10^{-3} (4.10×10^{-4})
Correlation	ρ	0.40 (0.07)		[0.08 (0.03); 0.14 (0.05)]
3. Second crop season at ESALQ				
Intercept	ξ_0	2.72 (0.02)	2.62 (9.96×10^{-3})	3.28 (0.05)
Ear diameter	ξ_1			
Ear length	ξ_2			0.02 (3.22×10^{-3})
Overdispersion	α	-0.04 (1.66×10^{-3})	-0.06 (2.90×10^{-4})	-0.02 (5.56×10^{-4})
Compound symmetry	d	0.01 (3.56×10^{-3})	3.18×10^{-3} (8.10×10^{-4})	1.72×10^{-3} (5.48×10^{-4})
Correlation	ρ	0.34 (0.06)	0.16 (0.04)	[0.14 (0.04); 0.21 (0.06)]
4. Second crop season at Sertãozinho				
Intercept	ξ_0	2.76 (0.02)		3.35 (0.05)
Ear diameter	ξ_1			
Ear length	ξ_2			0.01 (3.16×10^{-3})
Overdispersion	α	-0.03 (2.25×10^{-3})		-0.02 (5.19×10^{-4})
Compound symmetry	d	0.01 (3.49×10^{-3})		1.38×10^{-3} (4.55×10^{-4})
Correlation	ρ	0.28 (0.06)		[0.13 (0.04); 0.18 (0.05)]

Note that the occurrence of negative variance components is not an isolated phenomenon. In this application, the variance component related to the gamma distribution showed negative estimates for all characters across all environments.

In line with the calculated intraclass correlations and with the results presented in Section 4.3, positive correlations and no overdispersion implies negative α , given that ρ induces extra variance, which is removed by the negative estimate of the α component. Regarding the normal random effect, the estimates for the variance component d were positive due to positive intraclass correlations.

Although the PQL method has shown the best performance among the three estimation methods, it failed for the character kernel row number (KR) in First crop season at Sertãozinho and Second crop season at Sertãozinho. Ideally, additional methods should be developed for negative variance component estimation, especially when interest lies in inference of such effects.

4.6 Concluding Remarks

Hierarchical data are common in empirical research. For the analysis of continuous data, the linear mixed model is a flexible tool while the generalized linear mixed model is commonly used to model non-Gaussian data. Beyond inferences on the fixed effects, such models allow inferences about variance components. Although these quantities are positive when interpreted as variances, the occurrence of negative estimates is documented in the literature.

In this work, we investigated the variance components of the Poisson-Gamma-Normal

(PGN) model. This model, developed for count data, accommodates hierarchies as well as overdispersion in the data, through normal and gamma distributed random effects, respectively. The variance components associated with these distributions were studied theoretically and a maize breeding data were used for illustration purposes.

The application underscores that negative estimates of variance components can occur, especially when the variability is low and there is an expected negative intraclass correlation due to intra-specific competition.

References

BOOTH, J. G.; CASELLA, G.; FRIEDL, H.; HOBERT, J. P. Negative binomial loglinear mixed models. **Statistical Modelling**, London, v. 3, n. 3, p. 179-191, 2003

BRESLOW, N. E. Extra-poisson variation in log-linear models. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, London, v. 33, n. 1, p. 38-44, 1984.

BRESLOW, N. E.; CLAYTON, D. G. Approximate inference in generalized linear mixed models. **Journal of the American Statistical Association**, Boston, v. 88, n. 421, p. 9-25, 1993.

BRITTON, T. Tests to detect clustering of infected individuals within families. **Biometrics**, Washington, v. 53, p. 98-109, 1997.

CHERNOFF, H. On the distribution of the likelihood ratio. **Annals of Mathematical Statistics**, Ann Arbor, v. 25, p. 573-578, 1954.

KENNY, D. A.; MANNETTI, L.; PIERRO, A.; LIVI, S.; KASHY, D. A. The statistical analysis of data from small groups. **Journal of Personality and Social Psychology**, Boston, v. 83, n. 1, p. 126-137, 2002.

LAWLESS, J. F. Negative binomial and mixed poisson regression. **The Canadian Journal of Statistics**, Ottawa, v. 15, n. 3, p. 209-225, 1987.

MCCULLAGH, P.; NELDER, J. A. **Generalized linear models**. London: Chapman & Hall, 1989. 532 p.

MOLENBERGHS, G.; VERBEKE, G. **Models for discrete longitudinal data**. New York: Springer, 2005. 683 p.

MOLENBERGHS, G.; VERBEKE, G. A note on a hierarchical interpretation for negative variance components. **Statistical Modelling**, London, v. 11, n. 5, p. 389-408, 2011.

MOLENBERGHS, G.; VERBEKE, G.; DEMÉTRIO, C. G. B. An extended random-effects approach to modeling repeated, overdispersed count data. **Lifetime Data Analysis**, Hingham, v. 13, n. 4, p. 513-531, 2007.

NELDER, J.A. The interpretation of negative components of variance. **Biometrika**, Cambridge, v. 41, n. 3-4, p. 544-548, 1954.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. **Journal of the Royal Statistical Society: Series A**, London, v. 135, n. 3, p. 370-384, 1972.

PRYSELEY, A.; TCHONLAFI, C.; VERBEKE, G.; MOLENBERGHS, G. Estimating negative variance components from gaussian and non-gaussian data: A mixed models approach. **Computational Statistics & Data Analysis**, Amsterdam, v. 55, n. 2, p. 1071-1085, 2011.

VANGENEUGDEN, T.; MOLENBERGHS, G.; VERBEKE, G.; DEMÉTRIO, C. G. B. Marginal correlation from an extended random-effects model for repeated and overdispersed counts. **Journal of Applied Statistics**, Abingdon, v. 38, n. 2, p. 215-232, 2011.

VERBEKE, G.; MOLENBERGHS, G. **Linear mixed models for longitudinal data**. New York: Springer-Verlag, 2000. 568 p.

WELSH, A. H. **Aspects of Statistical Inference**. Hoboken: Wiley Series in Probability and Statistics, 2011. 480 p.

5 GENERAL CONCLUSIONS AND FUTURE RESEARCH

In this work, we have proposed solutions to statistical problems arising from agricultural sciences, especially genetics and breeding. For this, we used the combined models, extended generalized linear mixed models that allow for overdispersion. Such models address a cluster-induced correlation as well as overdispersion in models for non-Gaussian outcomes. This is done conveniently through the inclusion of conjugate random effects which allows the models estimation in standard software, through partial marginalization.

In Chapters 2 and 3 we proposed the calculation of the genetic contribution over the total variability of a phenotype, for time-to-event and count traits. Then, the important concept of heritability, in essence, was extended to non-Gaussian outcomes. The resulting expressions are relatively simple and SAS implementations were available. Next the addressed cases, binary outcomes are also common in agricultural and livestock studies. Such case can be encompassed in future work using the Bernoulli combined model.

For the Poisson case (Chapter 3), we showed that, when linear models are used for count outcomes, the genetic effect on the variability of the trait may be overestimated. So, based on such results, one would expect a response to selection that, in fact, may not happen. This is always a drawback in breeding programs, since the losses with no response to selection can be large. It is noteworthy that during the analysis we did not consider the lattice design structure and treated the data as coming from a completely randomized design. For illustration purposes this was an acceptable choice, but it is clear that further studies are needed to include the effects of complex designs into the proposed calculations of heritability. Likewise, the genotype-environment interaction has important implications and should be modeled in future studies.

Furthermore, in genetics and plant breeding it is often of interest study correlations between characters, which can be discrete or continuous. For instance, one may want to study the correlation between number of tassel branches and corn yield (ton/ha). Joint modeling framework can be explored and applied to get such correlations in future research.

In Chapter 4 we turned attention for occurrence and proper treatment of negative variance components in the Poisson combined model. Through our data analyses, we have shown that this phenomenon can be common in agricultural data and associated, for example, with negative intraclass correlations between individuals within the same cluster. Negative variance components may have important practical interpretations and their estimates can be failed with the estimation methods currently available. Therefore, additional methods should be developed for negative variance component estimation.

Still regarding this theme, there is a need for follow-up work involving general assessments through simulation studies and the investigation of the variance components in the Weibull

combined model. The latter can be relatively easy since the Weibull and Poisson models are, in essence, mostly similar.

Finally, all analysis were performed using SAS, but alternative implementations should be explored. A R package, for example, can be developed to handle the combined models and the calculation of the genetic attribute addressed in this work.

APPENDIX

APPENDIX A - Model Elements for the Weibull-Gamma-Normal and Weibull-Normal Models

From the marginal density of the combined model specified by (2.3)–(2.5), Molenberghs et al. (2010) derived the moment and mean expressions:

$$\begin{aligned} E(Y_{ij}^k) &= \frac{\alpha_j B(\alpha_j - k/\rho, k/\rho + 1)}{\lambda^{k/\rho} \beta_j^{k/\rho}} \exp \left(-\frac{k}{\rho} x'_{ij} \boldsymbol{\xi} + \frac{k^2}{2\rho^2} z'_{ij} D z_{ij} \right), \\ E(Y_{ij}) &= \frac{\alpha_j B(\alpha_j - 1/\rho, 1/\rho + 1)}{\lambda^{1/\rho} \beta_j^{1/\rho}} \exp \left(-\frac{1}{\rho} x'_{ij} \boldsymbol{\xi} + \frac{1}{2\rho^2} z'_{ij} D z_{ij} \right), \end{aligned}$$

which easily leads to the variance expression:

$$\begin{aligned} \text{Var}(Y_{ij}) &= \frac{\alpha_j}{\lambda^{2/\rho} \beta_j^{2/\rho}} \exp \left(-\frac{2}{\rho} x'_{ij} \boldsymbol{\xi} + \frac{1}{\rho^2} z'_{ij} D z_{ij} \right) \\ &\quad \times \left[B(\alpha_j - 2/\rho, 2/\rho + 1) \exp \left(\frac{1}{\rho^2} z'_{ij} D z_{ij} \right) \right. \\ &\quad \left. - \alpha_j B \left(\alpha_j - \frac{1}{\rho}; \frac{1}{\rho} + 1 \right)^2 \right]. \end{aligned} \quad (1)$$

The exponential version follows from setting $\rho = 1$ in the above expressions.

Now we will derive the moment, mean and variance expressions for the specific case of no overdispersion and only one normal random effect, i.e., the Weibull-Normal model (W-N). It can be easily obtained from the Weibull-based GLMM density as described in the following. From

$$f(\mathbf{y}_i | \mathbf{b}_i) = \prod_{j=1}^{n_i} \lambda \rho y_{ij}^{\rho-1} e^{x'_{ij} \boldsymbol{\xi} + z'_{ij} \mathbf{b}_i} e^{-\lambda y_{ij}^\rho e^{x'_{ij} \boldsymbol{\xi} + z'_{ij} \mathbf{b}_i}}$$

let us derive the moments by

$$E(Y_{ij}^k | \mathbf{b}_i) = \lambda \rho \int y_{ij}^{k+\rho-1} e^{x'_{ij} \boldsymbol{\xi} + z'_{ij} \mathbf{b}_i} e^{-\lambda y_{ij}^\rho e^{x'_{ij} \boldsymbol{\xi} + z'_{ij} \mathbf{b}_i}} dy_{ij},$$

i.e., the k th moment conditional upon $\mathbf{b}_i$. Then consider $\varphi_{ij} = \lambda \exp(x'_{ij} \boldsymbol{\xi} + z'_{ij} \mathbf{b}_i)$, from which we deduce

$$E(Y_{ij}^k | \mathbf{b}_i) = \rho \int y_{ij}^{k+\rho-1} \varphi_{ij} e^{-\varphi_{ij} y_{ij}^\rho} dy_{ij} = \int y_{ij}^k e^{-(\varphi_{ij} y_{ij}^\rho)} d(\varphi_{ij} y_{ij}^\rho).$$

Writing $t = \varphi_{ij} y_{ij}^\rho$ we have $y_{ij}^k = t^{k/\rho} \varphi_{ij}^{-k/\rho}$ and

$$E(Y_{ij}^k | \mathbf{b}_i) = \int t^{k/\rho} \varphi_{ij}^{-k/\rho} e^{-t} dt = \varphi_{ij}^{-k/\rho} \int_0^\infty t^{k/\rho} e^{-t} dt = \varphi_{ij}^{-k/\rho} \Gamma \left(\frac{k}{\rho} + 1 \right).$$

Then, the moment expression is

$$E(Y_{ij}^k) = \frac{\Gamma \left(\frac{k}{\rho} + 1 \right)}{\lambda^{k/\rho} e^{\mu_{ij} k/\rho}} e^{(k^2/2\rho^2) z'_{ij} D z_{ij}}, \quad (2)$$

where μ_{ij} is the fixed part of the linear predictor. From (2) we have the first and second moments:

$$\begin{aligned} E(Y_{ij}) &= \frac{\Gamma\left(\frac{1}{\rho} + 1\right)}{\lambda^{1/\rho}} \exp\left(-\frac{1}{\rho}x'_{ij}\boldsymbol{\xi} + \frac{1}{2\rho^2}z'_{ij}Dz_{ij}\right), \\ E(Y_{ij}^2) &= \frac{\Gamma\left(\frac{2}{\rho} + 1\right)}{\lambda^{2/\rho}} \exp\left(-\frac{2}{\rho}x'_{ij}\boldsymbol{\xi} + \frac{4}{2\rho^2}z'_{ij}Dz_{ij}\right), \end{aligned}$$

which easily lead to the variance expression:

$$\begin{aligned} \text{Var}(Y_{ij}) &= \frac{1}{\lambda^{2/\rho}} \exp\left(-\frac{2}{\rho}x'_{ij}\boldsymbol{\xi} + \frac{1}{\rho^2}z'_{ij}Dz_{ij}\right) \\ &\times \left[\Gamma\left(\frac{2}{\rho} + 1\right) \exp\left(-\frac{1}{\rho^2}z'_{ij}Dz_{ij}\right) - \Gamma\left(\frac{1}{\rho} + 1\right)^2 \right]. \end{aligned} \quad (3)$$

APPENDIX B - Closed-form Expressions for Heritability for the Time-to-Event Case

Expression (2.6) is a ratio of the variances of the combined model presented in (1), where the variance in the numerator is defined by setting $D = 0$, that is, no other source of variability is taken into account except than these predicted by the mean-variance link and due to overdispersion. Using $\beta_j = 1/\alpha_j$ and after some simplification we have:

$$\zeta_{W_{ij}} = \frac{B\left(\alpha_j - \frac{2}{\rho}; \frac{2}{\rho} + 1\right) - \alpha_j B\left(\alpha_j - \frac{1}{\rho}; \frac{1}{\rho} + 1\right)^2}{\exp\left(\frac{1}{\rho^2} z'_{ij} D z_{ij}\right) v(\alpha_j, \rho, D, z_{ij})}, \quad (4)$$

with

$$\begin{aligned} & v(\alpha_j, \rho, D, z_{ij}) \\ &= \left[B\left(\alpha_j - \frac{2}{\rho}; \frac{2}{\rho} + 1\right) \exp\left(\frac{1}{\rho^2} z'_{ij} D z_{ij}\right) - \alpha_j B\left(\alpha_j - \frac{1}{\rho}; \frac{1}{\rho} + 1\right)^2 \right]. \end{aligned}$$

Using the gamma-function expansion of the beta functions involved, (2.6) follows.

The heritability expression can be extended to a more general case where more than one normal random effect are present. This occurs, for example, when the genetic effect is supplemented with an environmental effect. For this, let us consider the combined model for overdispersion and correlated data with two normal random effects, $\mathbf{v}_i$ and $\mathbf{w}_i$. In this case the combined model can be expressed as:

$$\begin{aligned} & f(\mathbf{y}_i | \boldsymbol{\theta}_i, \mathbf{v}_i, \mathbf{w}_i) \\ &= \prod_{j=1}^{n_i} \rho \theta_{ij} y_{ij}^{\rho-1} e^{x'_{ij} \boldsymbol{\xi} + z_{1ij} \mathbf{v}_i + z_{2ij} \mathbf{w}_i} e^{-y_{ij}^{\rho} \theta_{ij} e^{x'_{ij} \boldsymbol{\xi} + z_{1ij} \mathbf{v}_i + z_{2ij} \mathbf{w}_i}}, \end{aligned} \quad (5)$$

$$\begin{aligned} f(\boldsymbol{\theta}_i) &= \prod_{j=1}^{n_i} \frac{1}{\beta_j^{\alpha_j} \Gamma(\alpha_j)} \theta_{ij}^{\alpha_j-1} e^{-\theta_{ij}/\beta_j}, \\ f(\mathbf{v}_i) &= \frac{1}{(2\pi)^{q/2} |D_1|^{1/2}} e^{-(1/2) \mathbf{v}_i' D_1^{-1} \mathbf{v}_i}, \end{aligned} \quad (6)$$

$$f(\mathbf{w}_i) = \frac{1}{(2\pi)^{q/2} |D_2|^{1/2}} e^{-(1/2) \mathbf{w}_i' D_2^{-1} \mathbf{w}_i} \quad (7)$$

and the variance equals:

$$\begin{aligned} \text{Var}(Y_{ij}) &= \frac{\alpha_j}{\lambda^{2/\rho} \beta_j^{2\rho}} \exp\left(\frac{-2}{\rho} x'_{ij} \boldsymbol{\xi} + \frac{1}{\rho^2} z'_{1ij} D_1 z_{1ij} + \frac{1}{\rho^2} z'_{2ij} D_2 z_{2ij}\right) \\ &\quad \times v(\alpha_j, \rho, D_1, D_2, z_{1ij}, z_{2ij}), \end{aligned} \quad (8)$$

with

$$\begin{aligned} & v(\alpha_j, \rho, D_1, D_2, z_{1ij}, z_{2ij}) \\ &= \left[B\left(\alpha_j - 2/\rho; 2/\rho + 1\right) \exp\left(\frac{1}{\rho^2} z'_{1ij} D_1 z_{1ij} + \frac{1}{\rho^2} z'_{2ij} D_2 z_{2ij}\right) \right. \\ &\quad \left. - \alpha_j B\left(\alpha_j - \frac{1}{\rho}; \frac{1}{\rho} + 1\right)^2 \right]. \end{aligned}$$

The contribution from model-induced measurement error, overdispersion, and from the random effect w_i over the total variability is:

$$\zeta_{W_{ij}} = \zeta_{N,W_{ij}} / \zeta_{D,W_{ij}}, \quad (9)$$

where

$$\begin{aligned} \zeta_{N,W_{ij}} = & B \left(\alpha_j - \frac{2}{\rho}; \frac{2}{\rho} + 1 \right) \exp \left(\frac{1}{\rho^2} z'_{2ij} D_2 z_{2ij} \right) \\ & - \alpha_j B \left(\alpha_j - \frac{1}{\rho}; \frac{1}{\rho} + 1 \right)^2 \end{aligned} \quad (10)$$

and

$$\zeta_{D,W_{ij}} = \exp \left(\frac{1}{\rho^2} z'_{1ij} D_1 z_{1ij} \right) v(\alpha_j, \rho, D_1, D_2, z_{1ij}, z_{2ij}), \quad (11)$$

with

$$\begin{aligned} v(\alpha_j, \rho, D_1, D_2, z_{1ij}, z_{2ij}) = & \left[B \left(\alpha_j - \frac{2}{\rho}; \frac{2}{\rho} + 1 \right) \exp \left(\frac{1}{\rho^2} z'_{1ij} D_1 z_{1ij} + \frac{1}{\rho^2} z'_{2ij} D_2 z_{2ij} \right) \right. \\ & \left. - \alpha_j B \left(\alpha_j - \frac{1}{\rho}; \frac{1}{\rho} + 1 \right)^2 \right]. \end{aligned}$$

Using the gamma function, we can rewrite

$$\begin{aligned} \zeta_{N,W_{ij}} = & \Gamma(\alpha_j) \Gamma \left(\alpha_j - \frac{2}{\rho} \right) \Gamma \left(\frac{2}{\rho} + 1 \right) \exp \left(\frac{1}{\rho^2} z'_{2ij} D_2 z_{2ij} \right) \\ & - \Gamma \left(\alpha_j - \frac{1}{\rho} \right)^2 \Gamma \left(\frac{1}{\rho} + 1 \right)^2 \end{aligned} \quad (12)$$

and

$$\zeta_{D,W_{ij}} = \exp \left(\frac{1}{\rho^2} z'_{1ij} D_1 z_{1ij} \right) v(\alpha_j, \rho, D_1, D_2, z_{1ij}, z_{2ij}), \quad (13)$$

with

$$\begin{aligned} v(\alpha_j, \rho, D_1, D_2, z_{1ij}, z_{2ij}) = & \left[\Gamma(\alpha_j) \Gamma \left(\alpha_j - \frac{2}{\rho} \right) \Gamma \left(\frac{2}{\rho} + 1 \right) \exp \left(\frac{1}{\rho^2} z'_{1ij} D_1 z_{1ij} + \frac{1}{\rho^2} z'_{2ij} D_2 z_{2ij} \right) \right. \\ & \left. - \Gamma \left(\alpha_j - \frac{1}{\rho} \right)^2 \Gamma \left(\frac{1}{\rho} + 1 \right)^2 \right]. \end{aligned}$$

Then, the heritability coefficient when there are two normal random effects is

$$H_{W_{ij}}^2 = 1 - \zeta_{W_{ij}}. \quad (14)$$

As before, setting $\rho = 1$ leads to the special case of an exponential time-to-event distribution. For this case, we have that

$$\zeta_{E_{ij}} = \zeta_{N,E_{ij}} / \zeta_{D,E_{ij}},$$

where

$$\zeta_{N,E_{ij}} = 2\Gamma(\alpha_j)\Gamma(\alpha_j - 2) \exp(z'_{2ij}D_2z_{2ij}) - \Gamma(\alpha_j - 1)^2$$

and

$$\begin{aligned} \zeta_{D,E_{ij}} = & \exp(z'_{1ij}D_1z_{1ij}) [2\Gamma(\alpha_j)\Gamma(\alpha_j - 2) \exp(z'_{1ij}D_1z_{1ij} + z'_{2ij}D_2z_{2ij}) \\ & - \Gamma(\alpha_j - 1)^2]. \end{aligned}$$

The heritability coefficient is

$$H^2_{E_{ij}} = 1 - \zeta_{E_{ij}}. \quad (15)$$

Specific cases arise when there is only one variance component in D_1 and in D_2 , say d_1 and d_2 . This leads to:

$$\zeta_{W_j} = \frac{\Gamma(\alpha_j)\Gamma\left(\alpha_j - \frac{2}{\rho}\right)\Gamma\left(\frac{2}{\rho} + 1\right)\exp\left(\frac{1}{\rho^2}d_2\right) - \Gamma\left(\alpha_j - \frac{1}{\rho}\right)^2\Gamma\left(\frac{1}{\rho} + 1\right)^2}{\exp\left(\frac{1}{\rho^2}d_1\right)v(\alpha_j, \rho, d_1, d_2)},$$

with

$$\begin{aligned} v(\alpha_j, \rho, d_1, d_2) = & \left[\Gamma(\alpha_j)\Gamma\left(\alpha_j - \frac{2}{\rho}\right)\Gamma\left(\frac{2}{\rho} + 1\right)\exp\left(\frac{1}{\rho^2}d_1 + \frac{1}{\rho^2}d_2\right) \right. \\ & \left. - \Gamma\left(\alpha_j - \frac{1}{\rho}\right)^2\Gamma\left(\frac{1}{\rho} + 1\right)^2 \right] \end{aligned}$$

in the Weibull case and leads to

$$\zeta_{E_j} = \frac{2\Gamma(\alpha_j)\Gamma(\alpha_j - 2)\exp(d_2) - \Gamma(\alpha_j - 1)^2}{\exp(d_1)[2\Gamma(\alpha_j)\Gamma(\alpha_j - 2)\exp(d_1 + d_2) - \Gamma(\alpha_j - 1)^2]}$$

in the exponential case.

APPENDIX C - SAS Implementation Related to Chapter 2

A SAS program, using the procedure NLMIXED, for the purely Exponential-Normal (E-N) model is as follows:

```
proc nlmixed data=d.nelorelp tech=NRRIDG qpoints=100
    maxit=50 corr;
    title 'Exponential-normal model';
    parms beta0=-2.9 beta1=-0.07 beta2=0.3;
    logkappa= beta0 + b + beta1*(sex=1) + beta2*(pby=96);
    kappa = exp(logkappa);
    ll = logkappa - t160_mont * kappa;
    model t160_mont ~ general(ll);
    random b ~ normal(0,d*d) subject=sire;
    estimate 'Variance' d*d;
run;
```

The special case of the purely Exponential simply follows from removing the RANDOM statement and the b random term from the linear predictor. These models also can be fitted using the GLIMMIX procedure, that allows for the Exponential distribution.

The Weibull-normal model was fitted in the NLMIXED procedure using:

```
proc nlmixed data=d.nelorelp tech=NRRIDG qpoints=50
    maxit=50 corr;
    title 'Weibull-normal model';
    parms beta0=-72 beta1=1.78 beta2=-0.33 rho=20;
    logkappa=b + beta0 + beta1*(sex=1) + beta2*(pby=96);
    kappa = exp(logkappa);
    ll = log (rho) + (rho-1)*log(t160_mont)
        + logkappa - t160_mont**rho * kappa;
    model t160_mont ~ general(ll);
    random b ~ normal(0,d*d) subject=sire;
    estimate 'Variance' d*d;
run;
```

and, similar to the Exponential case, the purely Weibull follows from removing the RANDOM statement and the random term of the linear predictor. The *general likelihood* or user-defined likelihood feature of this procedure is properly suited to implement the models with Gamma distributed effects, using the partially integrated version of the models. The code to implement the Weibull-Gamma model is:

```
proc nlmixed data=d.nelorelp qpoints=5 maxit=50 corr;
```

```

title 'Weibull-gamma model';
parms beta0=-72 beta1=1.78 beta2=-0.33 rho=20;
logkappa=beta0 + beta1*(sex=1) + beta2*(pby=96);
kappa = exp(logkappa);
ll = logkappa + log(rho) + (rho-1) * log(t160_mont) -
      (gamma+1) * log(1 + kappa * t160_mont**rho / gamma);
model t160_mont ~ general(ll);
run;

```

and the combined model, i.e., with Gamma and Normal random effects is:

```

proc nlmixed data=d.nelorelp qpoints=5 maxit=50 corr;
title 'Weibull-gamma-normal model';
parms beta0=-110 beta1=2.82 beta2=-0.65 rho=30;
logkappa=b + beta0 + beta1*(sex=1) + beta2*(pby=96);
kappa = exp(logkappa);
ll = logkappa + log(rho) + (rho-1) * log(t160_mont) -
      (gamma+1) * log(1 + kappa * t160_mont**rho / gamma);
model t160_mont ~ general(ll);
random b ~ normal(0,d*d) subject=sire;
estimate 'Variance' d*d;
run;

```


APPENDIX D - Model Elements for the Poisson-Gamma-Normal and Poisson-Normal Models

The mean and variance expressions for the (PGN), (3.1) and (4.1), were presented by Molenberghs et al. (2007). The mean vector $\boldsymbol{\mu}_i = E(\mathbf{Y}_i)$ has components

$$\mu_{ij} = \phi \exp \left(x'_{ij} \boldsymbol{\xi} + \frac{1}{2} z'_{ij} D z_{ij} \right) \quad (16)$$

and the variance-covariance matrix is given by

$$\text{Var}(\mathbf{Y}_i) = M_i + M_i(P_i - J_{n_i})M_i, \quad (17)$$

where M_i is a diagonal matrix with the vector $\boldsymbol{\mu}_i$ along the diagonal and the $(j, k)^{th}$ element of P_i equals

$$p_{i,jk} = \exp \left(\frac{1}{2} z'_{ij} D z_{ik} \right) \frac{\sigma_{i,jk} + \phi_{ij} \phi_{ik}}{\phi_{ij} \phi_{ik}} \exp \left(\frac{1}{2} z'_{ik} D z_{ij} \right).$$

Note that $\sigma_{i,jk}$ is the $(j, k)^{th}$ element of Σ_i .

Vangeneugden et al. (2011) presented closed-form expression for the correlation function for the general case of the combined model and its specific cases. Considering the combined model with arbitrary fixed- and random-effects structures, the variance, deriving from (17) equals:

$$\begin{aligned} \text{Var}(Y_{ij}) &= \phi_{ij} e^{x'_{ij} \boldsymbol{\xi} + \frac{1}{2} z'_{ij} D z_{ij}} + \sigma_{i,jj} e^{2x'_{ij} \boldsymbol{\xi} + 2z'_{ij} D z_{ij}} \\ &\quad + \phi_{ij}^2 e^{2x'_{ij} \boldsymbol{\xi} + 2z'_{ij} D z_{ij}} (e^{z'_{ij} D z_{ij}} - 1). \end{aligned}$$

Likewise, the covariance can be written as:

$$\begin{aligned} \text{Cov}(Y_{ij}, Y_{ik}) &= \phi_{ij} e^{x'_{ij} \boldsymbol{\xi} + \frac{1}{2} z'_{ij} D z_{ij}} \left[\left(\frac{\sigma_{i,jk}}{\phi_{ij} \phi_{ik}} + 1 \right) e^{\frac{1}{2} (z'_{ij} D z_{ik} + z'_{ik} D z_{ij})} - 1 \right] \times \\ &\quad \times \phi_{ik} e^{x'_{ik} \boldsymbol{\xi} + \frac{1}{2} z'_{ik} D z_{ik}}. \end{aligned}$$

The correlation between two measures j and k on the same cluster (subject) i then is:

$$\text{Corr}(Y_{ij}, Y_{ik}) = \frac{\text{Cov}(Y_{ij}, Y_{ik})}{\sqrt{\text{Var}(Y_{ij}) \text{Var}(Y_{ik})}}.$$

These expressions also produce their simplified counterparts for important special cases. For the (P-N), when only normal random effects are present, the mean vector components slightly simplify:

$$\mu_{ij} = \exp \left(x'_{ij} \boldsymbol{\xi} + \frac{1}{2} z'_{ij} D z_{ij} \right),$$

and the variance-covariance matrix is

$$\text{Var}(\mathbf{Y}_i) = M_i + M_i(e^{Z_i D Z'_i} - J_{n_i})M_i.$$

Similar logic as in the (PGN) leads to the correlation expression for this special case, considering:

$$\text{Var}(Y_{ij}) = e^{x'_{ij}\boldsymbol{\xi} + \frac{1}{2}Z'_{ij}DZ_{ij}} + e^{2x'_{ij}\boldsymbol{\xi} + Z'_{ij}DZ_{ij}}(e^{Z'_{ij}DZ_{ij}} - 1),$$

and

$$\text{Cov}(Y_{ij}, Y_{ik}) = e^{x'_{ij}\boldsymbol{\xi} + \frac{1}{2}Z'_{ij}DZ_{ij}} \left(e^{Z'_{ij}DZ_{ik}} - 1 \right) e^{x'_{ik}\boldsymbol{\xi} + \frac{1}{2}Z'_{ik}DZ_{ik}}.$$

APPENDIX E - SAS Implementation Related to Chapter 3

The SAS programs, using the procedure NLMIXED, for the linear mixed model and Poisson models are as follows.

```
proc nlmixed data=env1 corr;
title 'NTB: Linear mixed model';
parms beta0=16.36 sigma=3.23;
mean = beta0 + b;
model ntb ~ normal(mean,sigma*sigma);
random b ~ normal(0,d*d) subject=gen;
estimate 'Variance Prog.' d*d;
estimate 'Variance Residual' sigma*sigma;
run;
```

The special case of the Poisson model simply follows from removing the RANDOM statement and the b effect in the linear predictor.

```
proc nlmixed data=env1 tech=NRRIDG qpoints=50 corr;
title 'NTB: Poisson-Normal model';
parms beta0=2.7948;
eta = beta0 + b;
lambda = exp(eta);
model ntb ~ poisson(lambda);
random b ~ normal(0,d*d) subject=gen;
estimate 'Variance Prog.' d*d;
run;
```

The above program makes use of the built-in Poisson likelihood. The Poisson-Gamma-Normal model, or the combined model, need to be implemented using the so-called *general likelihood*, i.e., the user defined likelihood feature. The resulting program is as follows:

```
proc nlmixed data=env1 tech=NRRIDG qpoints=50;
title 'NTB: Poisson-Gamma-Normal model';
parms beta0=2.79 alpha=200 d=0.06775;
eta = beta0 + b;
lambda = exp(eta);
ll = lgamma(alpha + ntb) - lgamma(alpha) + ntb*log(1/alpha) -
      (ntb + alpha)*log(1 + (1/alpha)*lambda) + ntb*eta;
model ntb ~ general(ll);
random b ~ normal(0,d*d) subject=gen;
estimate 'Variance Prog.' d*d;
run;
```

The special case of the Poisson-Gamma model or, equivalently, the negative binomial model, follows from removing the RANDOM statement in the previous program.

APPENDIX F - Results for Kernel-row Number and Kernel Number per Row

The estimates of the linear mixed models and Poisson models for the traits kernel-row number and kernel number per row are displayed in Tables 1 and 2, respectively.

Table 1 – Maize study. Parameter estimates and standard errors for the regression coefficients in (1) the linear mixed model (LMM), (2) the purely Poisson model (P--), (3) the Poisson-Normal model (P-N), (4) the Poisson-Gamma model (PG-), and (5) the Poisson-Gamma-Normal model (PGN), considering the trait kernel-row number in the four environments

Effect	Par.	(LMM)	(P--)	(P-N)	(PG-)	(PGN)
1. First crop season at ESALQ						
Intercept	ξ_0	10.059 (0.939)	2.615 (0.012)	2.615 (0.012)	2.615 (0.012)	2.615 (0.012)
Ear diameter	ξ_1	0.739 (0.192)				
Gamma param.	α				9116.39 (11350)	1200.00 (539.77)
Var. of progenies	σ_g^2	0.107 (0.048)		1.15×10^{-14}		1.18×10^{-14}
Var. residual	σ_r^2	1.219 (0.082)				
-2log-likelihood		1518.6	2236.1	2236.1	-21636	-21632
AIC		1526.6	2238.1	2240.1	-21632	-21626
2. First crop season at Sertãozinho						
Intercept	ξ_0	8.505 (0.997)	2.156 (0.200)	2.156 (0.200)		
Ear diameter	ξ_1	1.096 (0.207)	0.097 (0.042)	0.097 (0.042)		
Gamma param.	α					
Var. of progenies	σ_g^2	0.260 (0.081)		6.51×10^{-27}		
Var. residual	σ_r^2	1.325(0.089)				
-2log-likelihood		1581.8	2245.1	2245.1		
AIC		1589.8	2249.1	2251.1		
3. Second crop season at ESALQ						
Intercept	ξ_0	12.236 (0.983)	2.615 (0.012)	2.615 (0.012)	2.615 (0.012)	2.615 (0.012)
Ear diameter	ξ_1	0.307 (0.209)				
Gamma param.	α				9131.74 (11593)	800.00 (300.44)
Var. of progenies	σ_g^2	0.534 (0.136)		3.42×10^{-17}		2.51×10^{-12}
Var. residual	σ_r^2	1.285 (0.087)				
-2log-likelihood		1593.9	2251.3	2251.3	-21646	-21640
AIC		1601.9	2253.3	2255.3	-21642	-21634
4. Second crop season at Sertãozinho						
Intercept	ξ_0	9.967 (1.064)	2.619 (0.012)	2.619 (0.012)	2.619 (0.012)	2.619 (0.012)
Ear diameter	ξ_1	0.779 (0.220)				
Gamma param.	α				8410.39 (10023)	1000.00 (414.27)
Var. of progenies	σ_g^2	0.248 (0.077)		1.39×10^{-16}		8.75×10^{-15}
Var. residual	σ_r^2	1.312 (0.088)				
-2log-likelihood		1575.6	2245.9	2245.9	-21772	-21767
AIC		1583.6	2247.9	2249.9	-21768	-21761

Table 2 – Maize study. Parameter estimates and standard errors for the regression coefficients in (1) the linear mixed model (LMM), (2) the purely Poisson model (P--), (3) the Poisson-Normal model (P-N), (4) the Poisson-Gamma model (PG-), and (5) the Poisson-Gamma-Normal model (PGN), considering the trait kernel number per row in the four environments

Effect	Par.	(LMM)	(P--)	(P-N)	(PG-)	(PGN)
1. First crop season at ESALQ						
Intercept	ξ_0	27.312 (2.247)	3.304 (0.090)	3.304 (0.090)		
Ear length	ξ_2	0.621 (0.128)	0.019 (0.005)	0.019 (0.005)		
Gamma param.	α					
Var. of progenies	σ_g^2	1.518 (0.580)		4.12×10^{-34}		
Var. residual	σ_r^2	12.632 (0.853)				
-2log-likelihood		2672.0	2868.5	2868.5		
AIC		2680.0	2872.5	2874.5		
2. First crop season at Sertãozinho						
Intercept	ξ_0	25.002 (2.014)	3.257 (0.080)	3.257 (0.080)	3.257 (0.082)	3.257 (0.082)
Ear length	ξ_2	0.779 (0.116)	0.023 (0.005)	0.023 (0.005)	0.023 (0.005)	0.023 (0.005)
Gamma param.	α				800.05 (213.48)	800.01 (213.47)
Var. of progenies	σ_g^2	1.456 (0.582)		2.66×10^{-27}		3.63×10^{-18}
Var. residual	σ_r^2	13.377 (0.903)				
-2log-likelihood		2697.5	2881.1	2881.1	-99866	-99866
AIC		2705.5	2885.1	2887.1	-99860	-99858
3. Second crop season at ESALQ						
Intercept	ξ_0	25.137 (1.938)	3.195 (0.080)	3.195 (0.080)		
Ear length	ξ_2	0.694 (0.120)	0.025 (0.005)	0.025 (0.005)		
Gamma param.	α					
Var. of progenies	σ_g^2	2.203 (0.696)		4.09×10^{-21}		
Var. residual	σ_r^2	11.106 (0.751)				
-2log-likelihood		2623.8	2840.6	2840.6		
AIC		2631.8	2844.6	2846.6		
4. Second crop season at Sertãozinho						
Intercept	ξ_0	27.636 (1.848)	3.300 (0.086)	3.300 (0.086)		
Ear length	ξ_2	0.490 (0.114)	0.017 (0.005)	0.017 (0.005)		
Gamma param.	α					
Var. of progenies	σ_g^2	1.710 (0.555)		2.53×10^{-31}		
Var. residual	σ_r^2	9.870 (0.665)				
-2log-likelihood		2561.7	2810.8	2810.8		
AIC		2569.7	2814.8	2816.8		

APPENDIX G - Precision Estimation of Correlation in the Poisson-Gamma-Normal Model

Standard errors for the intraclass correlations in the Poisson-Gamma-Normal (4.2) model were obtained using the delta method. We consider $\rho = \zeta_N/\zeta_D$, apply the delta method first to numerator and denominator, and then to the ratio. Assume W is the variance-covariance matrix of $\begin{pmatrix} \zeta_N \\ \zeta_D \end{pmatrix}$, then $\text{Var}(\zeta) \cong T'WT$, where

$$T = \frac{\partial \zeta}{\partial(\zeta_N, \zeta_D)} = \begin{pmatrix} 1/\zeta_D \\ -\zeta_N/\zeta_D^2 \end{pmatrix}.$$

To estimate W , we also use the delta method. At this stage, let ϕ be the parameter vector relevant for ζ_N and ζ_D and V be its variance-covariance matrix. Then, $W \cong S'VS$, with $S = \frac{\partial(\zeta_N, \zeta_D)}{\partial\phi}$.

The S matrix is

$$\begin{pmatrix} \frac{\partial \zeta_N}{\partial d} & \frac{\partial \zeta_D}{\partial d} \\ \frac{\partial \zeta_N}{\partial \alpha} & \frac{\partial \zeta_D}{\partial \alpha} \\ \frac{\partial \zeta_N}{\partial \xi_0} & \frac{\partial \zeta_D}{\partial \xi_0} \end{pmatrix},$$

where

$$\begin{aligned} \frac{\partial \zeta_N}{\partial d} &= e^{\xi_0} e^{\frac{1}{2}d} \left[\frac{1}{2}(e^d - 1) + e^d \right], \\ \frac{\partial \zeta_D}{\partial d} &= e^{\xi_0} e^{\frac{1}{2}d} \left[\frac{1}{2}(3e^d \alpha + 3e^d - 1) \right], \\ \frac{\partial \zeta_N}{\partial \alpha} &= 0, \\ \frac{\partial \zeta_D}{\partial \alpha} &= e^{\xi_0} e^{\frac{3}{2}d}, \\ \frac{\partial \zeta_N}{\partial \xi_0} &= e^{\xi_0} e^{\frac{1}{2}d} (e^d - 1), \\ \frac{\partial \zeta_D}{\partial \xi_0} &= e^{\xi_0} e^{\frac{1}{2}d} (e^d \alpha + e^d - 1). \end{aligned}$$

The correlation expression depend on estimates of fixed and random effects, that is, d , α and ξ_0 , and the V matrix should contain all their variances and covariances. However, the SAS procedure GLIMMIX provides a variance-covariance matrix for the random effects and another variance-covariance matrix for the fixed effects, separately. Then, the covariances between random and fixed estimates in the V matrix were set to zero, which are not very different from the true values.

When there are covariate effects, as for the trait number of grain rows, additional parameters, say ξ , need to be considered in the method. In such case and due to the linearity in the linear

predictor of GLMs, the derivatives will be the same as those with respect to ξ_0 , but depending on the covariate values.

APPENDIX H - Results From Other Estimation Methods in the Poisson-Gamma-Normal Model

In Tables (3) and (4) the estimates for the (PGN), considering Laplace and adaptive Gauss-Hermite quadrature as estimation methods, are displayed. Clearly, both methods showed poor performance for this application, where negative variance components are expected.

The Laplace method did not converge in some cases and failed to estimate the random effects and their standard errors. In a simulation study, Pryseley et al. (2011) investigated the performance of the PQL and Laplace methods in the face of negative variance components for binary clustered data, by means of generalized linear mixed models. In their study, Laplace approximations were more accurate than PQL and convergence was easier to reach, different than we noticed in this application for count data and using the (PGN).

Although quadrature methods are generally considered the most accurate ones, they adopt a hierarchical perspective and cannot be used when negative variance components are allowed for. Confirming this, the method failed in all situations and should not be used in such studies.

Table 3 – Maize study. Parameter estimates (standard errors) for the (PGN), considering random intercept case, for number of tassel branches (NTB), kernel row number (KR) and kernel number per row (KN). The estimation method used was Laplace

Effect	Par.	NTB	KR	KN
1. First crop season at ESALQ				
Intercept	ξ_0	2.79 (0.02)		3.30 (0.09)
Ear diameter	ξ_1			
Ear length	ξ_2			0.02 (5.13×10^{-3})
Overdispersion	α	1.09×10^{-7} (2.56×10^{-4})		7.41×10^{-9} (.)
Compound symmetry	d	4.59×10^{-3} (2.18×10^{-3})		-5.02×10^{-3} (.)
2. First crop season at Sertãozinho				
Intercept	ξ_0		2.16 (0.20)	3.26 (0.08)
Ear diameter	ξ_1		0.10 (0.04)	
Ear length	ξ_2			0.02 (4.60×10^{-3})
Overdispersion	α		5.83×10^{-9} (.)	5.12×10^{-9} (.)
Compound symmetry	d		-4.04×10^{-3} (.)	-5.25×10^{-3} (.)
3. Second crop season at ESALQ				
Intercept	ξ_0		2.62 (0.01)	3.20 (0.08)
Ear diameter	ξ_1			
Ear length	ξ_2			0.03 (4.94×10^{-3})
Overdispersion	α		2.96×10^{-9} (.)	1.37×10^{-9} (.)
Compound symmetry	d		-7.70×10^{-4} (.)	-3.26×10^{-3} (.)
4. Second crop season at Sertãozinho				
Intercept	ξ_0	2.76 (0.02)		3.30 (0.09)
Ear diameter	ξ_1			
Ear length	ξ_2			0.02 (5.32×10^{-3})
Overdispersion	α	3.98×10^{-8} (.)		8.37×10^{-9} (.)
Compound symmetry	d	9.99×10^{-3} (3.37×10^{-3})		-3.33×10^{-3} (.)

Table 4 – Maize study. Parameter estimates (standard errors) for the (PGN), considering random intercept case, for number of tassel branches (NTB), kernel row number (KR) and kernel number per row (KN). The estimation method used was adaptive Gauss-Hermite quadrature

Effect	Par.	NTB	KR	KN
1. First crop season at ESALQ				
Intercept	ξ_0	2.79 (0.02)	2.62 (0)	3.30 (0.09)
Ear diameter	ξ_1			
Ear length	ξ_2			0.02 (5.13×10^{-3})
Overdispersion	α	4.62×10^{-8} (.)	0 (.)	8.14×10^{-9} (.)
Compound symmetry	d	4.59×10^{-3} (2.18×10^{-3})	1.10×10^{-3} (1475.28)	2.27×10^{-20} (.)
2. First crop season at Sertãozinho				
Intercept	ξ_0	2.80 (0)	2.16 (0)	3.26 (0)
Ear diameter	ξ_1		0.10 (0)	
Ear length	ξ_2			0.02 (0)
Overdispersion	α	0 (8.20×10^{-5})	0 (4.60×10^{-5})	0 (3.20×10^{-5})
Compound symmetry	d	0 (.)	4.04×10^{-3} (.)	0 (.)
3. Second crop season at ESALQ				
Intercept	ξ_0	2.72 (0)	2.62 (0)	
Ear diameter	ξ_1			
Ear length	ξ_2			
Overdispersion	α	0 (.)	0 (4.60×10^{-5})	
Compound symmetry	d	0 (6.5×10^{-5})	7.70×10^{-4} (.)	
4. Second crop season at Sertãozinho				
Intercept	ξ_0	2.76 (0)	2.62 (0.01)	3.30 (0.09)
Ear diameter	ξ_1			
Ear length	ξ_2			3.30 (0.09)
Overdispersion	α	0 (6.10×10^{-5})	6.82×10^{-9} (.)	7.57×10^{-9} (.)
Compound symmetry	d	0.01 (.)	3.91×10^{-20} (.)	4.99×10^{-20} (.)