Quantifying the within-group contribution to the variability of count traits

Izabela R. C. Oliveira
ESALQ, University of São Paulo, Piracicaba, Brazil - izabela.rco@gmail.com

Geert Molenberghs
I-BioStat, Universiteit Hasselt, Hasselt, Belgium - geert.molenberghs@uhasselt.be

Clarice G.B. Demétrio*
ESALQ, University of São Paulo, Piracicaba, Brazil - clarice.demetrio@usp.br

Carlos T. S. Dias
ESALQ, University of São Paulo, Piracicaba, Brazil - ctsdias@usp.br

Cláudio L. Souza
ESALQ, University of São Paulo, Piracicaba, Brazil - clsouza@usp.br

Abstract: Heritability and repeatability are important concepts in animal and plant breeding and are quantified based on fitting a model to hierarchical data. When linear models can be used to fit to the data, these attributes are defined as ratios 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 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. The proposed methodology is illustrated using data from plant breeding programs. In the potato study, we have used our methodology to calculate the heritability for the count trait number of large tubers per plot. We considered two scenarios, with and without covariates. In the latter one, we showed how the combined model plays an important role in accommodating extra variability besides the intra-cluster correlation. In the tomato study, we have extended our approach to calculate the repeatability on trichomes count. 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 and repeatability change 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.

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

1 Introduction

Heritability, defined as the proportion of the genetic contribution over the total variability in a phenotype, 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. Repeatability is also an important concept in quantitative genetics and describes the proportion of phenotypic variance stemming from differences in repeated measures taken on the same individual. A measurement may be said to be repeatable when this proportion is relatively small.

When the outcomes are normally distributed, linear mixed models are frequently used to estimate the genetic and environmental effects by considering these factors as random terms in the model. In this case, heritability in the broad sense can be quantified as the ratio of the genotypic variance, σ_g^2 say, to the total phenotypic variance, σ_p^2. However, when the trait of interest 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. One often models such data using Poisson log-linear models. However, it has been observed recurrently that the mean-variance relationship for the Poisson model may not be met. In this paper, we use the combined models proposed by Molenberghs *et al.*(2010) for handling overdispersion (Hinde and Demétrio, 1998) and correlated data (Molenberghs and Verbeke, 2005) simultaneously, while obtaining heritability in the broad sense and repeatability based on count traits.

The paper is organized as follows. In Section 2, the motivating cases are described with analyses reported in Section 5. A review of the Poisson combined model for hierarchical and overdispersed count data is the subject of Section 3. The expressions to obtain heritability and repeatability coefficients for count traits are presented in Section 4.

2 Case Studies

2.1 The potato breeding study

A total of 31 clones of potato were evaluated, using an augmented blocks design, and several production traits were measured, including the number of large tubers, which are commercially interesting. The clones were replicated from two to six times to the experimental plots, each one consisting of about 10 plants. At 120 days after planting, all produced tubers were harvested and the number of large tubers per plot was counted.

2.2 Inheritance study of trichomes density in tomato

The epidermal outgrowths trichomes in tomato plants are related to resistance to whitefly, a plague of this crop. A completely randomized experiment, with one plant per experimental unit, was performed to study the inheritance of some types of trichomes in tomato, using plants from P_1, P_2, F_1, F_2, $BC_{1(1)}$ and $BC_{1(2)}$ populations. In each plant, three cuts were made and at each of them an area of 1 mm^2 was defined and the numbers of several different trichomes were counted, both in the abaxial and adaxial faces of the leaves. A main interest of this study lies in the repeatability calculation. In this work, we will consider only the glandular trichomes of types IV, VI, and VII counted in the adaxial face.

3 An Extended Poisson Model to Handle Hierarchical and Overdispersed Data

Combining ideas from the overdispersion models and the Poisson-normal model led Molenberghs, Verbeke and Demétrio (2007) and Molenberghs *et al.*(2010) to a model for repeated count data with overdispersion, assuming that $Y_{ij} \sim \text{Poi}(\lambda_{ij})$ with $\lambda_{ij} = \theta_{ij} \exp\left(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i\right)$, $\mathbf{b}_i \sim \text{N}(\mathbf{0}, \mathbf{D})$, $\text{E}(\boldsymbol{\theta}_i) = \text{E}[(\theta_{i1}, \ldots, \theta_{in_i})^T] = \Phi_i$ and $\text{Var}(\boldsymbol{\theta}_i) = \Sigma_i$. The $\boldsymbol{\mu}_i = \text{E}(\mathbf{Y}_i)$ has components: $\boldsymbol{\mu}_{ij} = \phi_{ij} \exp\left(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \frac{1}{2}\mathbf{z}_{ij}^T\mathbf{D}\mathbf{z}_{ij}\right)$, and the variance-covariance matrix $\text{Var}(\mathbf{Y}_i) = \mathbf{M}_i + \mathbf{M}_i\left(\mathbf{P}_i - \mathbf{J}_{n_i}\right)\mathbf{M}_i$, where $\mathbf{M}_i$ is a diagonal matrix with the vector $\boldsymbol{\mu}_i$ along the diagonal and the $(j,k)^{\text{th}}$ element of $\mathbf{P}_i$ equals $p_{i,jk} = \exp\left(\frac{1}{2}\mathbf{z}_{ij}^T\mathbf{D}\mathbf{z}_{ik}\right)\frac{\sigma_{i,jk}+\phi_{ij}\phi_{ik}}{\phi_{ij}\phi_{ik}}\exp\left(\frac{1}{2}\mathbf{z}_{ik}^T\mathbf{D}\mathbf{z}_{ij}\right)$. As special cases of this combined model we have the Poisson, Poisson-normal and negative-binomial models.

4 Derivation of Heritability for Count Data

Consider the Poisson-Gamma-Normal model and its variance. Also, without loss of generality, we set $\text{E}(\boldsymbol{\theta}_i) = 1$. Then $\text{Var}(Y_{ij}) = \mu_{ij} + \mu_{ij}(P_{i,jj} - 1)\mu_{ij}$, where $\mu_{ij} = \exp\left(x_{ij}^T\boldsymbol{\beta} + \frac{1}{2}\mathbf{z}_{ij}^T\mathbf{D}\mathbf{z}_{ij}\right) = \mu_{0ij}\mu_{1ij}$, and $P_{i,jj} =$

$\mu_{1ij}(\sigma_{i,jj}+1)\mu_{1ij} = \mu_{1ij}^2(\sigma_{i,jj}+1)$. The non-genetic contribution (the variance of the Poisson combined model, for $D=0$) over the total variability is:

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

The heritability, that is, the proportion of the total variability related to the genetic effect is:

$$H_{ij}^2 = 1-\xi_{ij}.$$

The repeatability is the contribution from overdispersion and between-individual variability over the total phenotypic variability in a population:

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

As before, this ratio places the variance of the Poisson combined model, for $D=0$, in the numerator and the full variance in the denominator. However, as the interest lies on repeatability, the normal random effect captures the correlation between measures within the same individual. Then, the denominator, which refers to the total phenotypic variance, includes the within-individual component as well. From (2) we can also estimate the within-individual contribution to the total variability, that is, $e_{wij} = 1-r_{ij}$.

The special case of no overdispersion follows easily for heritability and repeatability, by setting $\sigma_{i,jj}=0$.

5 Analysis of Case Studies

5.1 The potato breeding study

We considered the combined model and its special cases with linear predictor $\ln(\lambda_{ij}) = \beta_0 + b_i$, where β_0 is the effect common to all observations, b_i is the genetic random effect of the ith clone, assumed to be normally distributed with mean 0 and variance σ_g^2. In a second approach, we also evaluated the effect of the total weight of the plot as a covariate added to the linear predictor while modeling the number of large tubers.

When there are no covariates effects, the combined model fits better. The non-genetic contribution for the number of large tubers is $\xi_{ij} \cong 0.47$ and the heritability for this trait is $H_{ij}^2 \cong 0.53$, that is, about 53% of the total phenotypic variation is attributed to genetic variation among clones. When the covariate effect is considered, it captures some amount of variability that no longer needs to be modeled by the gamma random effect. Then, the Poisson-Normal model is the most appropriate. The non-genetic contribution over the total variability varied in the interval [0.50; 0.18] and the heritability values varied within the range $H^2 = [0.50; 0.82]$, depending on the covariate value.

5.2 Inheritance study of trichomes density in tomato

We considered the combined model and its special cases with linear predictor $\ln(\lambda_{ij}) = \beta_0 + b_i$, where β_0 is the overall effect and b_i is the random effect that captures the variability within the ith plant, assumed to be normally distributed with mean 0 and variance σ_w^2. The combined model is an improvement in fit relative to the other models. Thus, there are correlation and overdispersion effects to be modeled simultaneously. The repeatability is $r_{ij} \cong 9.39 \times 10^{-6}$ for the trichome type IV, $r_{ij} \cong 0.02$ for the trichome type VI and $r_{ij} \cong 0.36$ for the trichome type VII. In all cases, the repeatability values are very low, which indicate that the variability within individuals with respect to phenotypic variability is high.

6 Conclusions

In this paper, we have derived an expression for heritability, based on hierarchical count data, using Poisson-based mixed models. The focus was on the so-called combined model, which brings together a generalized

linear model for count data with both normal and gamma random effects, thus accommodating correlation between repeated measures and overdispersion. Importantly, as shown in Molenberghs *et al.*(2010), special cases of this combined model are the Poisson, Poisson-normal and negative-binomial models.

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 and repeatability coefficients are sufficiently simple and appealing, in particular in special cases.

In the potato study, we have used our methodology to calculate the heritability for the count trait number of large tubers per plot. We considered two scenarios, with and without covariates. In the latter one, we showed how the combined model plays an important role in accommodating extra variability besides the intra-cluster correlation. In the tomato study, we have extended our approach to calculate the repeatability on trichomes count.

In these models, heritability is a function rather than a constant. Practically, heritability and repeatability change 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.

Finally, a common interpretation of heritability is in terms of the coefficient of a regression parents-offspring (e.g., "regression of the value of a character measured for the son on the character measured for the father." This property of the heritability is clearly lost because of the non-normal and hence non-linear nature of our models. This is the price to pay for the use of a model that is more faithful to the data type being recorded.

Acknowledgments

Special Thanks to CAPES and CNPq.

References

Hinde, J. and Demétrio, C.G.B. (1998) Overdispersion: Models and estimation. *Computational Statistics and Data Analysis*, **27**, 151 – 170.

Molenberghs, G. and Verbeke, G. (2005) *Models for Discrete Longitudinal Data.* New York: Springer.

Molenberghs, G., Verbeke, G., and Demétrio, C.G.B. (2007) An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime Data Analysis* **13**, 513 – 531.

Molenberghs, G., Verbeke, G., Demétrio, C.G.B., Vieira, A. (2010) A Family of Generalized Linear Models for Repeated Measures With Normal and Conjugate Random Effects. *Statistical Science*, **25**, 325 – 347.