JOINT MODELING OF MEAN AND DISPERSION INCREASES THE ACCURACY OF RANDOM REGRESSION ON SIMULATED GROWTH CURVES

C.L. Nel[1,2] and A.R. Gilmour[3]

[1] Directorate: Animal Sciences: Western Cape Department of Agriculture, Elsenburg, 7607,
South Africa
[2]Department of Animal Sciences, Stellenbosch University, Matieland, 7602, South Africa
[3]11 Holman Way, Orange, NSW, 2800 Australia

SUMMARY

We considered the use of hierarchical generalized linear models (HGLM) to model a smooth trend for residual variance in a random regression analysis. The traditional RR model uses a step function to model the change in residual variance over time. We compared the approaches on their ability to recover input parameters from two simulated growth datasets with varying inputs of dispersion parameters. Solutions from the RR analysis using the step function were susceptible to changes in dispersion, while results from the HGLM analysis were stable and more accurate. The application to real growth datasets with more complicated features should be considered in the future.

INTRODUCTION

In random regression (RR) analysis on a longitudinal axis, say time, just as genetic variance changes smoothly along the axis, the residual variance/dispersion (σ_e^2) also changes smoothly with time. The latter is often approximated with using a step-function that divides the axis (e.g. age, days in milk) into intervals and estimating σ_e^2 within each interval. This easy solution does not model the dispersion smoothly, and the number of intervals, as well as the boundaries, is arbitrary.

In growth data of some species, the effect of scale with increasing age can lead to a sharp increase in dispersion during early stages of the s-shaped growth curve. Following RR results, it has been postulated that severe scale effects can dominate the analysis despite using a step function (Apiolaza *et al.* 2000; Nel *et al.* 2025). The issue is notable, since heterogeneity of residuals linked to factors in the mixed model can affect other solutions such as animal genetic effects (Hill 1984).

Joint modelling of the mean and dispersion was first proposed in the context of 'normal regression' (Aitkin 1987). In this setting, the computing procedure iterates between two generalized linear models (GLM): a model for the mean ($E(y)$) and a model for dispersion ($E(e_i^2) = \phi_i$), where ϕ_i is analysed as a gamma variable with a log-link function. More recently, this framework was extended to include random variables in the mean model as a hierarchical GLM (HGLM; Lee and Nelder 1996), and further extensions (e.g. double HGLM) have become of considerable interest in animal breeding (Rönnegård *et al.* 2010).

In this paper we evaluate the functionality of HGLM to model both dispersion and genetic variance smoothly in the context of RR along the longitudinal axis. It was proposed that the HGLM machinery could be beneficial compared to traditional RR models in enabling (1) a smooth trend for dispersion of residuals along the longitudinal trajectory, and (2) allowing the solutions to the mean model to be appropriately weighted according to the GLM estimates derived from the dispersion model. We compared the HGLM process to a traditional RR model based on the ability to recover parameters from a simulated growth dataset.

MATERIALS AND METHODS

Simulated data. All simulations were done using custom scripts in R. The design of the dataset roughly reflected that seen for an ostrich growth dataset (Nel *et al.* 2025) but augmented for simplicity. The simulated data consisted of a population recorded monthly for live weight (LW) as a longitudinal trait during the first nine months of age (t).

The population consisted of five discrete generations of 1000 progeny each. In each generation, 200 males and 200 females were randomly selected and assigned into pairwise breeding groups, each pair producing five full-sib progenies. The first generation was considered the base population with no records for LW. The final phenotype dataset thus consisted of 4000 individuals recorded for LW nine times. The pedigree contained 5000 identities, the progeny of 800 sires and dams.

For animal i recorded at age $j = 1, \dots, 9$, the phenotype for LW was simulated as: $y_{ikj} = \mu_j + cg_k + pe_i + a_{ij} + e_{ij}$where μ_j is the mean for age j following a third-degree polynomial curve (Figure 1) and cg_k is a contemporary group effects of $k = 1, \dots, 80$ levels (detail not shown). The animal permanent environmental (PE) variance (σ_{pe}^2) was constant across ages : $pe_i \sim N(0, \sigma_{pe}^2)$ but both the additive genetic variance $a_{ij} \sim N(0, \sigma_{aj}^2)$ and residual variance $e_{ij} \sim N\left(0, \sigma_{ej}^2\right)$ varied with age. The structure of this heterogeneity was introduced as follows:

Additive genetic effects were defined for each individual as linear (first order) reaction norms (RN) formulated to include both scale and rank type changes (Falconer 1990) across the age trajectory. This was achieved indirectly by first sampling breeding values at the start ($j = 1$; $\sigma_{a1}^2 = 5$) and at the end ($j = 9$; $\sigma_{a9}^2 = 15$) with correlation of 0.5 between breeding values (a_1,a_9) and calculating RN components intercept (a_0) and slope (a_x) after standardising j to values between -1 and 1.

The residual variances were defined two ways. In Sim_I, the dispersion of residuals was as defined $\phi_j = \exp\,(-0.52 + 1.73j - 0.13j^2)$, a second-degree polynomial trend that first showed a sharp increase in residual variance, but declined as the mean growth rate plateaus. In Sim_II, the curve $\phi_j = \exp\,(1.21 - 0.21j + 0.07j^2)$ was designed to reflect a sharp increase in dispersion of variance with age (Figure 1).

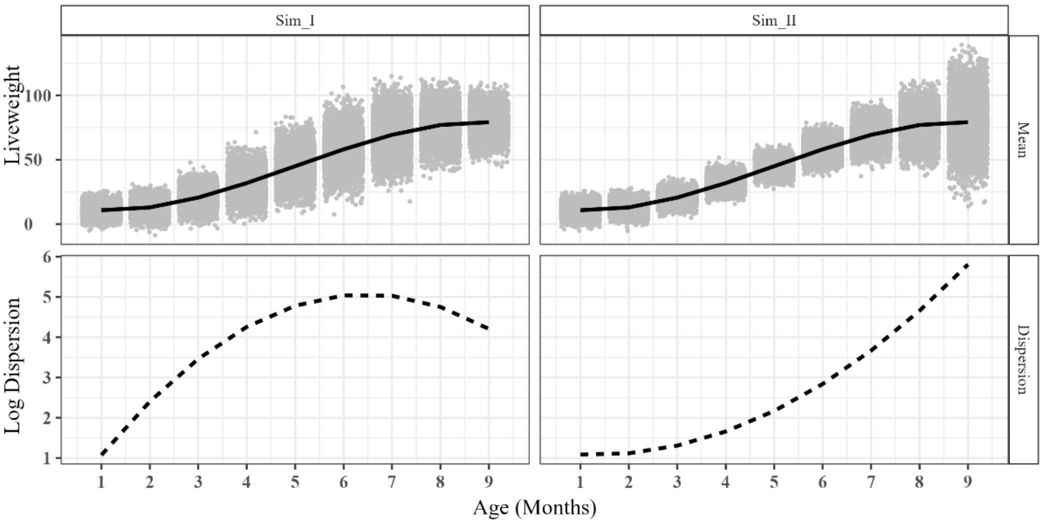

Figure 1. The distribution of LW phenotypes (gray) surrounding the mean growth trend (solid, black) as determined by the respective dispersion inputs (dashed, black) to simulations I and II

Statistical analysis. The data was analysed using the ASREML V4.2 software (Gilmour *et al.* 2021). For the traditional RR analysis, the residual was modelled as a step function in three intervals

(1-3, 4-6, 7-9 months). This design deliberately allowed an extent of change in $\phi_i$within intervals, a notable artefact of the data in Nel *et al.* (2025), but most likely true for the majority of previous RR analyses on growth data using a step function. The HGLM procedure uses the same mean model but weighted by inverse variances predicted by a dispersion model fitted simultaneously. In both cases, the linear model terms for the mean model were specified according to the simulation model design: $y = Xb + Za_0 + Za_x + Zpe + e$ assuming a normal distribution with an identity link but with heterogenous time dependent residual variance. Model testing was not a deliberate part of the study. The dispersion model in the HGLM analysis analysed the squared residuals from the mean model ($E(e_i^2) = \phi_i$) assuming a gamma distribution with a log link: $\log(\phi) = Xb$ where this Xb models a curvilinear (quadratic polynomial) trend. The appropriate covariance functions were used to find solutions for genetic variance (σ^2_{aj}) and accuracy of prediction ($cor(\widehat{ebv}, ebv)$) at selected age points, j. For each scenario, the simulation and analysis were repeated 20 times, and a mean of these results was used as the final estimate of either RR or HGLM.

RESULTS AND DISCUSSION

Mean and dispersion modelling. Both analyses showed reasonable convergence behaviour for both datasets (Sim_I and II), although RR usually converged in fewer iterations compared to HGLM. For HGLM, the dispersion model predicted both simulation $\log(\phi)$ trends with a very high accuracy with errors too small for it to be visually discernible from the true curves shown in Figure 1. The only example we found of HGLM applied to predict a smooth residual in RR was a dispersion model for lactation based on days in milk (Jaffrezic *et al.* 2000). To our knowledge, it has not been applied to growth curves, where the proportional change in the dispersion parameter ϕ is most likely of much greater magnitude than that expected for lactation data.

Variance parameters. It was apparent that RR analysis was susceptible to changes in dispersion, which varied depending on $\log(\phi)$ (Table 1). For Sim_I, with high dispersion at intermediate ages, the RR analysis was prone to overestimating genetic variance at the RN intercept component ($\sigma^2_{a_0}$) and underestimate variance of the RN slope ($\sigma^2_{a_x}$) – the latter by a considerable margin. This would explain the overestimation of the genetic correlation between breeding values at the ends of the trajectory ($p(a_1, a_9) = 0.85 > 0.5$; Table 1), which shows a partial failure in capturing reranking effects (Falconer 1990). In Sim_II, RR overestimated $\sigma^2_{a_x}$, likely due to high dispersion at the final age points. In both scenarios, the HLGM process was less affected by the curves of $\log(\phi)$, and solved the mean model with estimates very close to the true values – but perhaps also slightly underestimating the animal PE effect (Table 1).

Table 1. True values and solutions for variance parameters following RR and HGLM analysis. *The estimated values are the mean following 20 simulations

Parameters	True Values	Sim_I*		Sim_II*	
		RR	HGLM	RR	HGLM
σ^2_{pe}	10	9.02	9.73	9.63	9.64
σ^2_{a0}	7.15	8.64	7.28	8.00	7.31
σ^2_{ax}	1.89	0.88	1.75	2.59	1.88
$p(a_0, a_x)$	0.56	0.80	0.55	0.60	0.54
$p(a_1, a_9)$	0.50	0.83	0.54	0.43	0.51

Genetic variance at different ages and accuracy of prediction. The RR analysis estimated $\sigma^2_{a_j}$ of Sim_I close to the real values for first ($j = 1$) and last ($j = 9$) age points, but slightly higher values at intermediate ages (Table 2). For Sim_II, it overestimated $\sigma^2_{a_j}$ for 9 months of age by a notable margin, an indication of overestimating scale effects. HGLM results better aligned with the known input values which also resulted in a higher accuracy of prediction, most particularly for Sim_I (Table 2). Generally observing lower accuracy values at the ends of the trajectory was expected, since errors in estimating RN slope components are accentuated as distance to the intercept increases.

The computation in HGLM includes iterative updates of w_i, the i^{th} diagonal element of the weight matrix $W$of the mixed model equations, by the inverse of solutions ϕ_i from the dispersion model. In the current design, a benefit of this smooth adjustment from the HGLM process was clear, and, as argued by Rönnegård *et al.* (2010), the possibly important aspects of leverage should not be overlooked in animal breeding analysis with high heterogeneity. It should be noted, however, that the highly accurate modelling of $\log(\phi)$ would have been very important to the correct adjustment of W, and the process of model testing was ignored in this study.

Also, in real datasets, $\log(\phi)$ trends are unlikely to be well specified by inflexible polynomials. In a separate analysis, dispersion models making use of cubic smoothing splines also converged with accurate solutions. This is expected to be the most powerful application of HGLM to real data where both the magnitude and the change in $\log(\phi)$ are unknown.

Table 2. True values and solutions for genetic parameters and prediction accuracy following RR and HGLM analysis. *The estimated values are the mean following 20 simulations

Age	True Values	Genetic Parameters				Accuracy of EBVs			
(j)		Sim_I*		Sim_II*		Sim_I*		Sim_II*	
	$\sigma^2_{a_j}$	RR	HGLM	RR	HGLM	RR	HGLM	RR	HGLM
1	4.98	4.58	5.12	5.23	5.20	0.63	0.69	0.70	0.70
3	5.36	6.28	5.55	5.65	5.55	0.70	0.72	0.74	0.74
5	7.16	8.64	7.28	8.00	7.31	0.71	0.72	0.75	0.76
7	10.38	11.67	10.32	12.30	10.48	0.68	0.71	0.75	0.77
9	15.03	15.36	14.67	18.54	15.07	0.66	0.69	0.74	0.76

CONCLUSION

In the presence of sharp changes in dispersion, the genetic solutions from RR analysis can be affected by the choice of the step function. Analysis within the HGLM framework, in turn, can have more accurate genetic solutions as found here. However, the design of the simulated dataset was highly simplistic, and the process needs to be tested on real data.

REFERENCES

Aitkin M. (1987) *J. R. Stat. Soc. C-App.* **36**: 332.

Apiolaza L.A., Gilmour A.R. and Garrick D.J. (2000) *Can. J. Forest. Res.* **30**: 645.

Falconer D.S. (1990) *Gen. Res.* **56**: 57.

Gilmour A.R., Gogel B.J., Cullis B.R., Welham S.J. and Thompson A.N. (2021) ASReml User Guide Release 4.2 Functional Specification. www.vsni.co.uk.

Hill W.G. (1984) *Anim. Sci.* **39**: 473.

Jaffrezic F., White I.M.S., Thompson R. and Hill W.G. (2000) *J. Dairy Sci.* **83**: 1089.

Lee Y. and Nelder J.A. (1996) *J. R. Stat. Soc. B.* **58**: 619.

Nel C.L., Gilmour A.R., Muvhali P.T., Cloete S.W.P., Kekana M. and Engelbrecht A.E. (2025) *Brit. Poultry. Sci.* (Online Early).

Rönnegård L., Felleki M., Fikse F., Mulder H.A. and Strandberg E. (2010) *Genet. Sel. Evol.* **42**: 8.