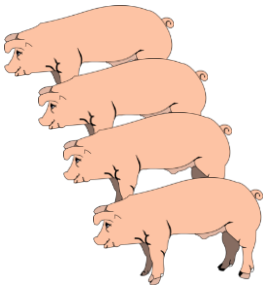
Estimation of genetic variance in Ve

Han Mulder



WAGENINGEN
UNIVERSITY & RESEARCH

Contents

- Data structures to estimate genetic variance in Ve
- Double hierarchical generalized linear model (DHGLM)
- Practical
  - Se of varav
  - DHGLM in asreml

WAGENINGEN
UNIVERSITY & RESEARCH

Learning outcomes

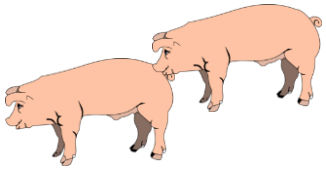
- To understand required data structures to estimate genetic variance in Ve
- To apply and interpret results of the double hierarchical generalized linear model

WAGENINGEN UNIVERSITY & RESEARCH

3

Required data and data structures

- Aim: estimate genetic differences in Ve
- Available data
  - Variance/standard deviation per animal
    - E.g. within-individual variance of repeated observation
    - Within-litter variance of piglet birth weight considered as trait of the sow
  - Variance/standard deviation per family
    - Half-sib families
    - Full-sib families
    - Clones



WAGENINGEN UNIVERSITY & RESEARCH

4

Standard error genetic variance in Ve

- Differences in within-family variance assuming additive model, e.g. paternal half-sibs
- $se(\sigma^2_{av,add}) = \sqrt{\frac{2/a^2\left[\left(\frac{2varW^2}{N+1}+a\sigma^2_{av,add}\right)^2+48\frac{varW^2}{(n-1)(n+1)^2}\right]}{N-1}}$
- $varW = (1-t)\sigma^2_P = (1-ah^2)\sigma^2_P$

WAGENINGEN UNIVERSITY & RESEARCH

Hill, 2004; J. Ind. Soc. Agric. Stat. 57:49-63.

Standard error genetic variance in Ve

- Exponential model
- $z_i = \log(\sum(X_{ij}-\bar{X}_i)^2/(n-1)$
- z_i has approximately a normal distribution with:
- $var(z_i) = \frac{2}{n-1} + \gamma^2$
- $\gamma^2 = CV^2 = a\sigma^2_{av,exp}\left(\frac{\sigma^2_E}{(1-t)\sigma^2_P}\right)^2$
- $se(\gamma^2) \cong \frac{\sqrt{8/m}}{n}$

t=intraclass correlation=ah^2

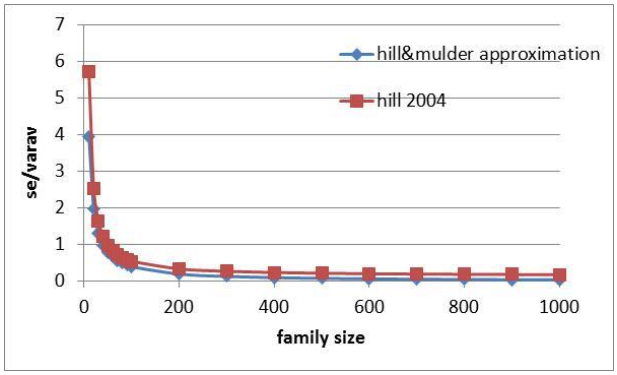
- $se(\sigma^2_{av,exp}) \cong se(\gamma^2) * \frac{1}{a} * \left(\frac{(1-t)\sigma^2_P}{\sigma^2_E}\right)^2$

m=number of families
n=family size

WAGENINGEN UNIVERSITY & RESEARCH

Hill and Mulder, 2010. Genet. Res. 92:381-395.

Standard error genetic variance in Ve

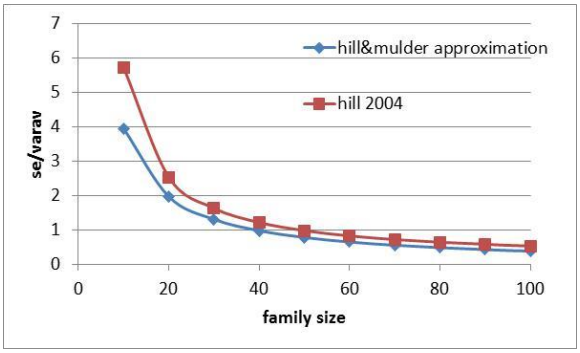


100 Half-sib families
Heritability=0.3
Varav,exp=0.05

Key message:
- Hill and Mulder approximation gives underestimation compared to Hill 2004 and simulations, but is in essence a bit simpler

WAGENINGEN UNIVERSITY & RESEARCH

Standard error genetic variance in Ve

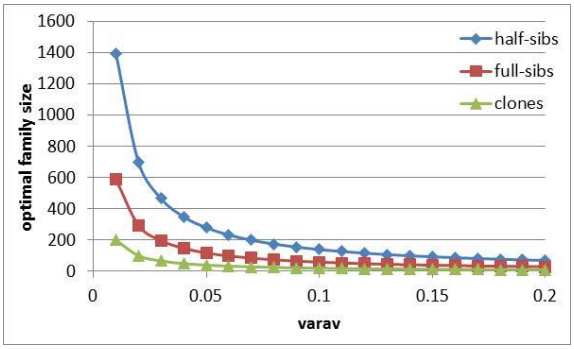


Key message:
- You need at least 100 offspring per family
- Large data sets needed

WAGENINGEN UNIVERSITY & RESEARCH

Optimum family size

- Given a fixed number of records
- Optimum family size: $2/\gamma^2$



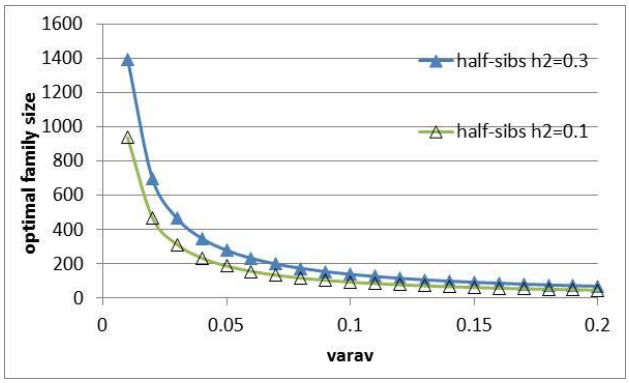
Heritability is 0.3

	σ^2_{av}=0.05	σ^2_{av}=0.10
Half-sibs	279	140
Full-sibs	118	59
Clones	40	20

Clones are ideal!

9

Optimum family size: effect of heritability



Key message

- Large family size needed 100-200 half-sib offspring
- For traits with low heritability, smaller family sizes are required

WAGENINGEN UNIVERSITY & RESEARCH

10

Estimating QTL/marker effects for Ve

- If marker affects the whole phenotypic variance
- $R^2 \cong p(1-p)(\lambda-1)^2$

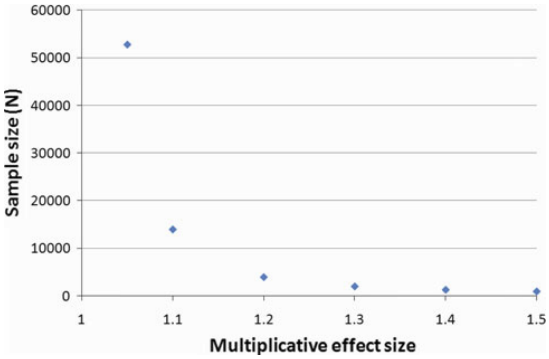
- R^2: amount of phenotypic variance explained by marker
- λ^x = multiplication factor phenotypic variance of genotype x=0, 1 or 2

- If marker affects only environmental variance:
- $R^2 \cong p(1-p)(\lambda-1)^2\left(\frac{\sigma_E^2}{\sigma_P^2}\right)^2$

WAGENINGEN UNIVERSITY & RESEARCH

Visscher and Posthuma, 2010. Behavior Genetics 40(5):728-733 11

Sample size unrelated individuals



Allele frequency=0.5
Type I error = 10^{-6}
Power=80%

Designs with 10,000 – 20,000 individuals needed to pick up QTL for variance

WAGENINGEN UNIVERSITY & RESEARCH

Visscher and Posthuma, 2010. Behavior Genetics 40(5):728-733. 12

Different models to estimate genetic variance in Ve

- Analysis of variance estimates per animal/family
  - Relatively simple
  - Rowe et al. (2006; Genet. Sel. Evol. 38:617-635)
  - Sell-Kubiak et al. (2015; J. Anim Sci. 93:900-911)
- Using squared residuals as response variables
  - Iterative REML method (Mulder et al., 2009; Animal 3:1673-1680)
  - **Double hierarchical generalized linear model** (Ronnegard et al., 2010; Genet. Sel. Evol. 42:8)
- Bayesian analysis
  - Sorensen and Waagepetersen (2003; Genet. Res. 82:207-222)

13

Analysis of variance estimates per animal/family

- Let's look at analysis of log(variance) of within-litter birth weight in pigs
  - Use of log(variance) gives estimates at level of exponential model
  - Estimates are comparable to DHGLM results
- Complexity
  - Need to account for heterogeneity of residual variance due to sample size
    - $var(var(x)) = \frac{2var^2(x)}{n+2}$

WAGENINGEN UNIVERSITY & RESEARCH

14

Analysis of variance estimates per animal/family

- Advantage
  - Simple
  - Intuitive
  - Use of standard packages
- Drawbacks
  - No feedback to correct for heterogeneity of residual variance in model for phenotype
  - Different number of observations for phenotype and variance

WAGENINGEN UNIVERSITY & RESEARCH

15

Bayesian analysis of Ve

- Prior distributions for Av and other effects on Ve
- MCMC implementation
- Special software needed GSEVM v2 (Ibanez-Eschriche et al. 2010, Journal of Animal Breeding and Genetics 127(3):249-251)
- Drawback: large computing time e.g. for large datasets

WAGENINGEN UNIVERSITY & RESEARCH

16

Comparison DHGLM and Bayesian method

Table 2. *Estimates and 95 % confidence intervals of chosen parameters for pigs litter size data in Model III (first section) and Model IV (second section) used by Sorensen & Waagepetersen (2003). Results obtained by Sorensen & Waagepetersen (2003) (first row in each section), by Rönnegård* et al. *(2010) (second row) and using IRWLS (third row)*

	Mean model		Residual variance model					
	Variances		Fixed effects*			Variances		Cor ρ
	σ_a^2	σ_p^2	$\beta_{\beta 0}$	$\beta_{d_{ins}}$	$\beta_{d_{par}}$	$\sigma_{a_d}^2$	$\sigma_{p_d}^2$	
Sorensen & Waagepetersen (2003) III	1·58	0·60	1·78	−0·16	0·34	0·11		−0·57
	1·13, 2·00	0·31, 0·96	1·65, 1·90	−0·24, −0·09	0·25, 0·43	0·08, 0·15		−0·72, −0·41
Rönnegård *et al.* (2010)	1·35	0·53	1·73	−0·17	0·32	0·13		
	0·99, 1·71	0·25, 0·81	1·61, 1·85	−0·23, −0·11	0·26, 0·39	0·09, 0·16		
IRWLS	1·61	0·34	1·70	−0·17	0·32	0·18		−0·49
	1·24, 1·97	0·08, 0·61	1·57, 1·82	−0·23, −0·11	0·26, 0·39	0·14, 0·22		−0·63, −0·36
Sorensen & Waagepetersen (2003) IV	1·62	0·60	1·77	−0·17	0·35	0·09	0·06	−0·62
	1·20, 2·05	0·30, 0·92	1·65, 1·89	−0·25, −0·09	0·26, 0·44	0·06, 0·13	0·05, 0·09	−0·80, −0·43
Rönnegård *et al.* (2010)	1·35	0·44	1·72	−0·17	0·32	0·09	0·06	
	1·00, 1·70	0·17, 0·71	1·62, 1·83	−0·23, −0·11	0·26, 0·39	0·05, 0·14	0·02, 0·11	
IRWLS	1·61	0·28	1·69	−0·17	0·32	0·15	0·05	−0·52
	1·25, 1·96	0·02, 0·54	1·57, 1·81	−0·23, −0·11	0·26, 0·39	0·10, 0·20	0·00, 0·09	−0·66, −0·37

*$\beta_{\beta 0}$ is the intercept term in the model for the residual variance, $\beta_{d_{ins}}$ is the fixed effect for insemination and $\beta_{d_{par}}$ is the fixed effect for the difference in first and second parity.

Parameters in the same direction, but not equal
Relative small data set ~10,000 litter size observations from ~4100 sows

WAGENINGEN UNIVERSITY & RESEARCH

Felleki et al. 2012; Genet. Res. 94:307-317.

17

Summary

- In general large experiments/datasets required to estimate genetic variance in Ve
- Simple methods can be appropriate, but have some drawbacks
- Bayesian hierarchical models and DHGLM models are more complex, but have more flexibility

WAGENINGEN UNIVERSITY & RESEARCH

18

DHGLM

DHGLM model in detail

- The model as used in Ronnegard et al. 2010
- The extension in Felleki et al. 2012
- Some background
- Effect of transformations
- Extensions of DHGLM

The DHGLM model

- Fixed effects on phenotype and residual variance; random genetic effects on phenotype and residual variance
- $\begin{bmatrix} \mathbf{y} \\ \mathbf{y_v} \end{bmatrix} = \begin{bmatrix} \mathbf{X} & \mathbf{0} \\ \mathbf{0} & \mathbf{X_v} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{b_v} \end{bmatrix} + \begin{bmatrix} \mathbf{Z} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z_v} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{u_v} \end{bmatrix} + \begin{bmatrix} \mathbf{e} \\ \mathbf{e_v} \end{bmatrix}$
- $var(e) = \phi$
- $\log(\phi) = \mathbf{X_v b_v + Z_v u_v}$

WAGENINGEN UNIVERSITY & RESEARCH

Rönnegård, L. et al. 2010. Genet. Sel. Evol. 42:8.

21

The DHGLM model – only fixed effects

- Maximum likelihood estimates for variance can be obtained by using Gamma GLM with squared residuals as response variable
- If fixed effects on mean are known without uncertainty
- $e_i^2 \sim \phi_i \chi_1^2$
- $E(e_i^2) = \phi_i$
- $Var(e_i^2) = 2\phi_i^2$
- Therefore: squared residual can be fitted using GLM with log link function together with gamma distribution for residual variance
- Note: Chi-square distribution is special case of gamma distribution

WAGENINGEN UNIVERSITY & RESEARCH

Rönnegård, L. et al. 2010. Genet. Sel. Evol. 42:8.

22

The DHGLM model – only fixed effects

- Fixed effects are estimated and we have only predicted residuals
- $E(e_i^2) \neq \phi_i$:
- variance of predicted residuals is smaller than the true variance, see example later.

- Therefore REML adjustment needed:
- $E(e_i^2/(1-h_i)) = \phi_i$
- And use weights for residual variance:
- $Var(e_i^2/(1-h_i)) = 2\phi_i^2/(1-h_i)$
- h_i = leverage of observation i

WAGENINGEN UNIVERSITY & RESEARCH

Rönnegård, L. et al. 2010. Genet. Sel. Evol. 42:8.

23

What is a leverage?

- Leverage: how much influence each data value y has on each predicted y ($\hat{y}$)
- $\mathbf{X'Xb} = \mathbf{X'y}$
- $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y}$

- $\hat{\mathbf{y}} = \mathbf{Xb} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'y} = \mathbf{Hy}$
- $\mathbf{H} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$ = The hat matrix

WAGENINGEN UNIVERSITY & RESEARCH

Hoaglin, D. C. and R. E. Welsh. 1978. Am. Stat. 32:17-22.

24

A simple genetic example

- Suppose we estimate a breeding value based on a single phenotype
- $EBV = h^2(P - \bar{P})$
- The residual: $\hat{e} = (P - \bar{P}) - EBV = (1 - h^2)(P - \bar{P})$
- $var(\hat{e}) = var\big((1 - h^2)(P - \bar{P})\big) = (1 - h^2)^2\sigma_P^2$

- Suppose $\sigma_P^2 = 1$, $h^2 = 0.3$, σ_E^2
- $var(\hat{e}) = 0.7^2 = 0.49$

- $var(\hat{e})/(1 - h) = \sigma_E^2 = 0.7$: h = leverage=heritability in this case; not in others
- Warning: h is not square root of heritability

WAGENINGEN UNIVERSITY & RESEARCH

25

DHGLM - Ronnegard

- Two univariate models implemented in ASREML
  - Updating of both models
  - Penalized quasi likelihood – approximation for h-likelihood method
  - No genetic correlation between phenotype and Ve
- Algorithm in ASREML
  1. Initialize W=I
  2. Estimate parameters of model y
  3. Calculate $y_{v,i} = \frac{e_i^2}{1-h_i}$ and $\mathbf{W_v} = diag(\frac{1-h}{2})$
  4. Estimate parameters for $y_{v,i}$
  5. Update $\mathbf{W} = diag(\hat{y}_{v,i})^{-1}$ based on step 4
  6. Iterate steps 2-5 until convergence of parameters

WAGENINGEN UNIVERSITY & RESEARCH

Rönnegård, L. et al. 2010. Genet. Sel. Evol. 42:8.

26

Felleki adjustment

- Bivariate linear mixed model
  - Estimate correlations between random effects on phenotype and its Ve
- Linearization of $y_{v,i} = \frac{e_i^2}{1-h_i}$; no need to use log-link function
- $y_{v,i} = \frac{e_i^2}{1-h_i} \cong \log\phi_i + \frac{\frac{e_i^2}{1-h_i} - \phi_i}{\phi_i}$
- First-order Taylor series approximation; equivalent to use of log-link function

WAGENINGEN UNIVERSITY & RESEARCH

Felleki et al., 2012; Genet. Res. 94:307-317.

27

DHGLM – IRWLS algorithm

- Implementation in ASREML
  - Iterative reweighted least squares approximation of h-likelihood
- Algorithm in ASREML
  1. Run model on $\mathbf{y}$ with homogeneous residual variance
  2. Calculate $\mathbf{y_v}$, $\mathbf{W}$, $\mathbf{W_v}$, where $\mathbf{W} = 1/\sigma_e^2$ in iteration 1
  3. Run bivariate model on $\mathbf{y}$ and $\mathbf{y_v}$.
  4. Update $\mathbf{y_v}$, $\mathbf{W}$, $\mathbf{W_v}$ based on output in 3
  5. Iterate steps 3-4 until convergence of parameters

WAGENINGEN UNIVERSITY & RESEARCH

Felleki et al., 2012; Genet. Res. 94:307-317.

28

```
!WORKSPACE 1800 !NOGRAPHICS !DEBUG !LOGFILE !RENAME !ARGS  1 2  // !DOPART  $1
DHGLM model of birth weight
    animal 32450 !I
    litter 2129  !I
    parity 10    !I
    sex 2        !I
    farm 15      !I
    ys 22        !I
    sow          !P
    bw           !M -99
    surv         !M -99
    Gval !=bw !-1.19 !*V10
    Ywt !=1. Gwt !=1.
Ped.txt
phenotype3.txt !maxit 1000 !skip 1 !DOPART $1

!PART 1 # normal model
bw ~ mu parity sex farm.ys !r sow litter
residual units

!Part 2
!ASUV !EXTRA 100 !SLOW
# in odd iterations, we use the predicted weights for the primary response
!IF ODD !CALC W1=EXP(R2-Y2) #redefine weights for Y1
!IF EVEN !CALC S1=1./W1; H0=MIN(H1/S1, .9999); Z2=MAX(R1*R1,.0001)/(1-H0)
!IF EVEN !CALC Y2=LOG(S1)+(Z2-S1)/S1 #redefine Y2
!IF EVEN !CALC W2=(1-H0)/2 #redefine weights for Y2

!ASSIGN gen 0.016 0.005 0.05
!ASSIGN lit 0.015 0.0 0.08

bw Gval !Weight Ywt !WT Gwt  ~ Trait Trait.parity Trait.sex Trait.farm.ys !r us(Trait,$gen).sow
us(Trait,$lit).litter !f mv
residual units.us(Trait) !VARIANCESCALE
```

29

- !Part 2
- !ASUV !EXTRA 100 !SLOW
- !IF ODD !CALC W1=EXP(R2-Y2) #redefine weights for Y1
  - W1: weight for residual variance trait 1 = phenotype
  - $\mathbf{W}1 = 1/\exp(\hat{y}_v)$: reciprocal of residual variance
- !IF EVEN !CALC S1=1./W1; H0=MIN(H1/S1, .9999); Z2=MAX(R1*R1,.0001)/(1-H0)
  - S1= residual variance per observation
  - H0=$\frac{h_{asreml}}{residual\ variance}$: h_{asreml}=in yhat file of ASREML
  - Z2=$\frac{e_i^2}{1-h_i}$
- !IF EVEN !CALC Y2=LOG(S1)+(Z2-S1)/S1 #redefine Y2
  - Y2: response variable $\mathbf{y_v}$
  - $y_{v,i} = \log\phi_i + \frac{\frac{e_i^2}{1-h_i} - \phi_i}{\phi_i}$
- !IF EVEN !CALC W2=(1-H0)/2 #redefine weights for Y2
  - W2: weight for residual variance trait 2 = $\mathbf{y_v}$
  - $\mathbf{W}2 = (1-h)/2$

30

Background of h-likelihood

- H-likelihood is statistical framework for statistical inference on models, suitable for hierarchical models
  - Similar to maximum likelihood or REML
  - Used when dispersion/variance is modeled
  - Very useful when modeling heterogeneity of residual variance
- Special packages needed such as hglm in R
  - Only small datasets and no pedigree data

WAGENINGEN UNIVERSITY & RESEARCH

Ronnegard, L., Shen, X. & Alam, M. (2010). HGLM: a package for fitting hierarchical generalized linear models. The R Journal 2, 20–28.

31

Model testing

- Adjusted Profile H-likelihood
- $APHL = 2LogL - \sum w_{v,i}\, e_{v,i}^2 - \sum \ln(\frac{1}{w_{v,i}})$
- Can be used together with likelihood ratio test or with AIC
- $AIC = APHL + 2t$
- t is number of variances/covariances
- Unfortunately not implemented yet in ASREML4, but one could calculate it based on yhat-file

WAGENINGEN UNIVERSITY & RESEARCH

Felleki et al., Genet. Res. 94:307-317; Mulder et al. 2013. Genet. Sel. Evol. 45:23

32

Bias in estimated variance components with animal models and single observations per animal

- Using animal model with single observations per animal give biased variance components

Table 2 Variance components of the classical linear mean animal model and two double hierarchical generalized linear models DHGLM1 and DHGLM2 (untransformed data)

Sub-model	Variance component	Classical	DHGLM1	DHGLM2
Mean	Sire-dam	-	-	0.071 ± 0.010
	Genetic[1]	0.303 ± 0.041	0.099 ± 0.023	0.283 ± 0.039
	Common environment	0.013 ± 0.008	0.063 ± 0.010	0.016 ± 0.008
	Residual	0.521 ± 0.022	-	-
	Heritability	0.362 ± 0.043	-	-
Variance	Sire-dam	-	0.051 ± 0.008	0.043 ± 0.008
	Genetic[1]	-	0.204 ± 0.033	0.174 ± 0.031

[1]Genetic variance is defined as 4*sire-dam variance.

WAGENINGEN UNIVERSITY & RESEARCH

Sonesson et al. 2013. Genet. Sel. Evol. 45:41.

33

Reasons for bias

- Too high dependency between EBV mean and residual
- The extremer observations would get a larger estimate for vEBV
- This would lower the 'heritability' for that particular observation, therefore lower EBV
- Therefore:
  - Genetic variance in phenotype too low
  - Genetic variance in residual variance too high

WAGENINGEN UNIVERSITY & RESEARCH

34

Solution to avoid bias

- Traits with repeated observations
- Traits with single observations
  - Sire models
  - Sire-dam models

WAGENINGEN UNIVERSITY & RESEARCH

Scaling of data

- Scaling: higher mean, higher variance
- Question:
  - Are we picking up scaling effects?
- This would mean a high positive genetic correlation between mean and variance

WAGENINGEN UNIVERSITY & RESEARCH

Evidence for scaling in fish

- Body weight of rainbow trout
- Genetic correlation between mean and Ve is:
  - Freshwater:0.30
  - Seawater: 0.79
- Genetic correlation between two environments
- Body weight: 0.70
- Ve of body weight: 0.56

WAGENINGEN UNIVERSITY & RESEARCH

Sae-Lim et al., 2015. Genet. Sel. Evol. 47:46.

37

What to do with scaling?

- Scaling can be removed by log transformation
- Removes heterogeneity of variance
- Removes the mean-variance relationship

WAGENINGEN UNIVERSITY & RESEARCH

38

Transformation rainbow trout example

- Effect of log-transformation
- Genetic correlation between mean and Ve is:
  - Freshwater: -0.83
  - Seawater: -0.62
- Genetic correlation between Ve in the two environments was -0.08
- Conclusion: log-transforming data has a large effect on genetic correlations. Too large?

WAGENINGEN UNIVERSITY & RESEARCH

Other transformations

- Square root transformation
- Cube-root transformation
- Box-Cox transformation
  - The optimal transformation to get distribution back to normal
  - $yt = \frac{y^{\lambda}-1}{\lambda}$
- Problem: after (box-cox) transformation, genetic correlations between mean and variance swap sign, but may be used to check whether genetic variance in Ve is not artefact of non-normality

WAGENINGEN UNIVERSITY & RESEARCH

Yang, Y., O. F. Christensen, and D. Sorensen. 2011. Genet. Res. 93:33-46.

Some extensions of DHGLM

- Estimating macro- and micro-environmental sensitivity
  - Reaction norm model for GxE combined with the DHGLM for differences in Ve
  - Mulder et al., 2013; Genet. Sel. Evol. 45:23.
- Estimating GxE for Ve
  - Fish in fresh water and sea water
  - Sae-Lim et al., 2015; Genet. Sel. Evol. 47:46.
- Estimating purebred-crossbred genetic correlation for Ve
  - Egg color in purebred and crossbred laying hens
  - Mulder et al. 2016; GSE 48:39

WAGENINGEN UNIVERSITY & RESEARCH

41

Some extensions of DHGLM

- Using genomic relationship matrix
  - Mulder et al. 2013; J. Dairy Sci. 96:7306-7317.
  - Sell-Kubiak et al. 2015; J. Anim. Sci. 93:1471-1480.
- Estimating relationships between Ve and other traits, such as fitness traits
  - Mulder et al. 2015; Genetics 199:1255-1269.
- Using DHGLM for genomic selection or GWAS
  - Ronnegard and Valdar, 2012; Genetics 188:435-447 and BMC Genet. 13:63
  - Ronnegard and Lee 2013 J. Anim. Breed. Genet.130:415-416.

WAGENINGEN UNIVERSITY & RESEARCH

42

Summary

- DHGLM is a relatively fast and good method to estimate genetic variance in Ve
  - Opportunities for extension
- Repeated observations needed per genotype
  - Single obs/animal: use sire or sire-dam model
  - Multiple obs/animal: use animal model
- Transformations may be useful to check whether genetic variance in Ve is not artefact of non-normality

WAGENINGEN UNIVERSITY & RESEARCH

43