A coefficient of determination (R^2) for generalized linear mixed models

Hans-Peter Piepho

Biostatistics Unit
Universität Hohenheim
Germany

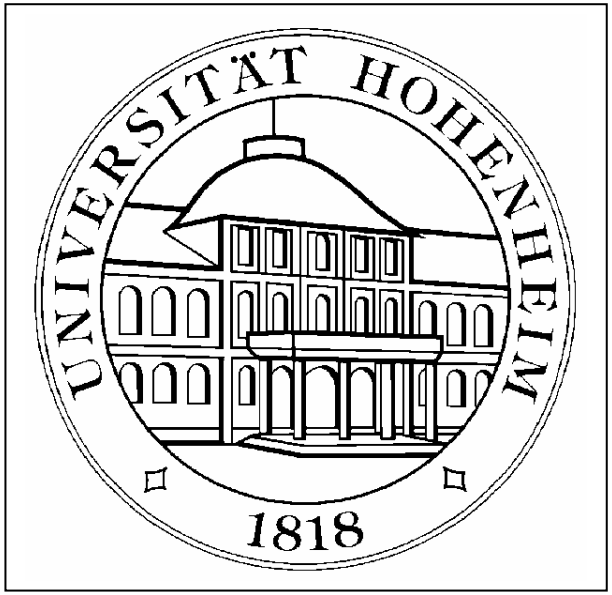

Table of contents

1. Coefficient of determination for linear model (LM)

Full model

$$y = X\beta + e\,, \tag{1}$$

where

y = response vector of length n

β = fixed effects vector

X = design matrix, and

$e \sim N\left(0, V = I_n \sigma_e^2\right)$ = residual error vector

1. Coefficient of determination for linear model (LM)

Null model

$$y = 1_n \phi + e \ , \qquad (2)$$

where

1_n = a vector of ones

ϕ = intercept

$e \sim N\left(0, V_0 = I_n \sigma_{e0}^2\right)$ = residual error vector

1. Coefficient of determination for linear model (LM)

The standard procedure

Error sum of squares for **full model**:

$$SS_{error}^{full} = y^T P_\beta y \quad \text{where} \quad P_\beta = I_n - X\left(X^T X\right)^- X^T$$

Error sum of squares for **null model**:

$$SS_{error}^{null} = y^T P_\phi y \quad \text{where} \quad P_\phi = I_n - n^{-1} 1_n 1_n^T$$

1. Coefficient of determination for linear model (LM)

Coefficient of determination (LM)

$$R^2 = 1 - \frac{SS_{error}^{full}}{SS_{error}^{null}}$$

$$R_{adj}^2 = 1 - \frac{(n-1)SS_{error}^{full}}{(n-p)SS_{error}^{null}} \quad \text{where} \quad p = rank(X)$$

1. Coefficient of determination for linear model (LM)

Coefficient of determination (LM)

$$R^2 = 1 - \frac{n^{-1} SS_{error}^{full}}{n^{-1} SS_{error}^{null}} = 1 - \frac{\hat{\sigma}_{e(ML)}^2}{\hat{\sigma}_{e0(ML)}^2}$$

$$R_{adj}^2 = 1 - \frac{(n-p)^{-1} SS_{error}^{full}}{(n-1)^{-1} SS_{error}^{null}} = 1 - \frac{\hat{\sigma}_{e(REML)}^2}{\hat{\sigma}_{e0(REML)}^2}$$

1. Coefficient of determination for linear model (LM)

What does R^2 estimate?

$$\Omega_\beta = \frac{\Delta\theta(V, V_0)}{\theta(V_0)} \quad , \tag{3}$$

where

$\theta(V)$ = total variance implied by the variance-covariance structure V

$$\Delta\theta(V, V_0) = \theta(V_0) - \theta(V)$$

= variance explained by effects added in full model relative to null model

1. Coefficient of determination for linear model (LM)

For LM

$\theta(V_0) = \sigma_{e0}^2$,

$\theta(V) = \sigma_e^2$, and

$\Delta\theta(V, V_0) = \sigma_{e0}^2 - \sigma_e^2$ and hence

$$\Omega_\beta = \frac{\sigma_{e0}^2 - \sigma_e^2}{\sigma_{e0}^2} = 1 - \frac{\sigma_e^2}{\sigma_{e0}^2} \qquad (4)$$

1. Coefficient of determination for linear model (LM)

Can we also get an R^2 for (generalized) linear mixed models?

Most common answers:

(1) No we can't!

(2) Yes, I have a specialized solution

1. Coefficient of determination for linear model (LM)

Extensions of R^2

Generalized linear models (GLM):
Zhang (2017)

Linear mixed models (LMM):
Edwards et al. (2008), Liu et al. (2008), Demidenko et al. (2012)

Generalized linear mixed models (GLMM):
Nagakawa and Schielzeth (2013), Jaeger et al. (2017, 2018),
Nakagawa et al. (2017), Stoffel et al. (2017)

$\Rightarrow$ No time to review in detail
$\Rightarrow$ None of these seemed general enough & easy to communicate

1. Coefficient of determination for linear model (LM)

Most prominent example:

Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2013, **4**, 133–142 doi: 10.1111/j.2041-210x.2012.00261.x

A general and simple method for obtaining R^2 from generalized linear mixed-effects models

Shinichi Nakagawa[1,2]* and Holger Schielzeth[3]

[1]National Centre for Growth and Development, Department of Zoology, University of Otago, 340 Great King Street, Dunedin 9054, New Zealand; [2]Department of Behavioral Ecology and Evolutionary Genetics, Max Planck Institute for Ornithology, Eberhard-Gwinner-Straße, 82319 Seewiesen, Germany; and [3]Department of Evolutionary Biology, Bielefeld University, Morgenbreede 45, 33615, Bielefeld, Germany

› 4900 citations and counting on SCOPUS !

2. Linear mixed models (LMM)

$$y = X\beta + Zu + e\ , \tag{5}$$

Z = design matrix

$u \sim N(0, G)$

$e \sim N(0, R)$

$y \sim N(X\beta, V)$ with

$$V = ZGZ^T + R \tag{6}$$

2. Linear mixed models (LMM)

Coefficient of determination for fixed effects in LMM

$$\Omega_\beta = \frac{\Delta\theta(V, V_0)}{\theta(V_0)}$$

V_0 is for the null model, in which $X\beta$ in (5) is replaced by $1_n\phi$.

2. Linear mixed models (LMM)

Requirements for definition of $\theta(V)$

$\Rightarrow$ allow for heterogeneity of variance

$\Rightarrow$ allow for covariance between observations

$\Rightarrow$ reduce to common R^2 for LM when random effects dropped and $R = I_n \sigma_e^2$

$\Rightarrow$ should be additive, i.e.,

$$\theta(V_1 + V_2) = \theta(V_1) + \theta(V_2) \tag{7}$$

2. Linear mixed models (LMM)

Marginal variance (mv)

$$mv(y_i) = v_{ii} \quad , \tag{8}$$

y_i = *i*-th element of y

v_{ij} = *ij*-th element of V

Average marginal variance (AMV)

$$\theta^{AMV}(V) = \frac{1}{n}\sum_{i=1}^{n} mv(y_i) = \frac{1}{n} trace(V) \tag{9}$$

2. Linear mixed models (LMM)

Average marginal variance (AMV)

$$\theta^{AMV}(V) = \frac{1}{n}\sum_{i=1}^{n} mv(y_i) = \frac{1}{n} trace(V)$$

The trace of a variance-covariance matrix is a common measure of total variance in multivariate analysis.

The major downside of this criterion is that it does not account for covariances $v_{ij} \quad (i \neq j)$ (Mustonen, 1997; Johnson and Wichern, 2002, p.139).

2. Linear mixed models (LMM)

"semivariance" (*sv*)

$$sv(y_i, y_j) = \frac{1}{2}\text{var}(y_i - y_j) = \frac{1}{2}(v_{ii} + v_{jj}) - v_{ij} \qquad (10)$$

Average semivariance (ASV)

$$\theta^{ASV}(V) = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j>i+1}^{n} sv(y_i, y_j) = \frac{1}{n-1} trace(VP_\phi) \qquad (11)$$

where $P_\phi = I_n - n^{-1}J_n$

(Webster and Oliver, 2007; Piepho, 2019)

2. Linear mixed models (LMM)

Average semivariance (ASV)

$$\theta^{ASV}(V) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j>i+1}^{n} sv(y_i, y_j) = \frac{1}{n-1} trace(VP_\phi) \qquad (11)$$

It is readily verified that

$$\theta^{AMV}(V) = \theta^{ASV}(V) = \sigma_e^2$$

when $V = R = I_n \sigma_e^2$ (i.e., for an LM) as required.

2. Linear mixed models (LMM)

Further motivation of ASV

(1) Total variance over a region in spatial statistics

The variance of $Z(\mathbf{x})$ within a region R of area $|R|$ is the double integral of the variogram:

$$\sigma_R^2 = \overline{\gamma}(R,R) = \frac{1}{|R|^2}\int_R \int_R \gamma(\mathbf{x}-\mathbf{x}')\,d\mathbf{x}d\mathbf{x}', \qquad (4.23)$$

(Webster and Oliver, 2007, p.61)

(2) Heritability

$$H^2 = \sigma_g^2 \Big/ \left(\sigma_g^2 + r^{-1}\sigma_e^2\right) = \sigma_g^2 \Big/ \left(\sigma_g^2 + \theta^{ASV}\left[V(\hat{\mu})\right]\right)$$

(Piepho and Möhring, 2007)

(3) A-efficiency of blocked designs

$$E_A = 2/(r \times apv) = \sigma_e^2 / \theta^{ASV}\left[V(\hat{\mu})\right]$$

(John & Williams, 1995)

4. Coefficient of determination for random effects

$$\Omega_u = \frac{\theta\left(ZGZ^T\right)}{\theta\left(V_0\right)} \tag{12}$$

Variance explained jointly by fixed and random effects:

$$\Omega_{\beta u} = \Omega_\beta + \Omega_u = 1 - \frac{\theta\left(R\right)}{\theta\left(V_0\right)} \tag{13}$$

4. Example LMM

Example 1: Yield trends for long-term variety trial data

$$y_{ijk} = \mu + G_i + L_j + Y_k + (LY)_{jk} + (GL)_{ij} + (GY)_{ik} + (GLY)_{ijk} \qquad (14)$$

y_{ijk} = mean yield of the *i*-th genotype in the *j*-th location and *k*-th year

μ = overall mean

G_i = main effect of the *i*-th genotype

L_j = main effect of the *j*-th location

Y_k = main effect of the *k*-th year

$(LY)_{jk}$ = *jk*-th location × year interaction

$(GL)_{ij}$ = *ij*-th genotype × location interaction

$(GY)_{ik}$ = *ik*-th genotype × year interaction

$(GLY)_{ijk}$ = residual comprising both genotype × location × year interaction as well as the error of a mean

4. Example LMM

Mackay et al. (2011)

- Take G_i and Y_k as fixed (can't take random because of time trend)
- All other effects random (i.i.d. normal with constant variance)
- Adjusted means for G_i assess genetic trend

  $\Rightarrow$ Plotted against year in which variety entered trial
- Adjusted means for Y_k assess non-genetic trend

  $\Rightarrow$ Plotted against calendar year
- Estimate trend by linear regression based on adjusted means for G_i and Y_k
- Look at one intensity at a time

4. Example LMM

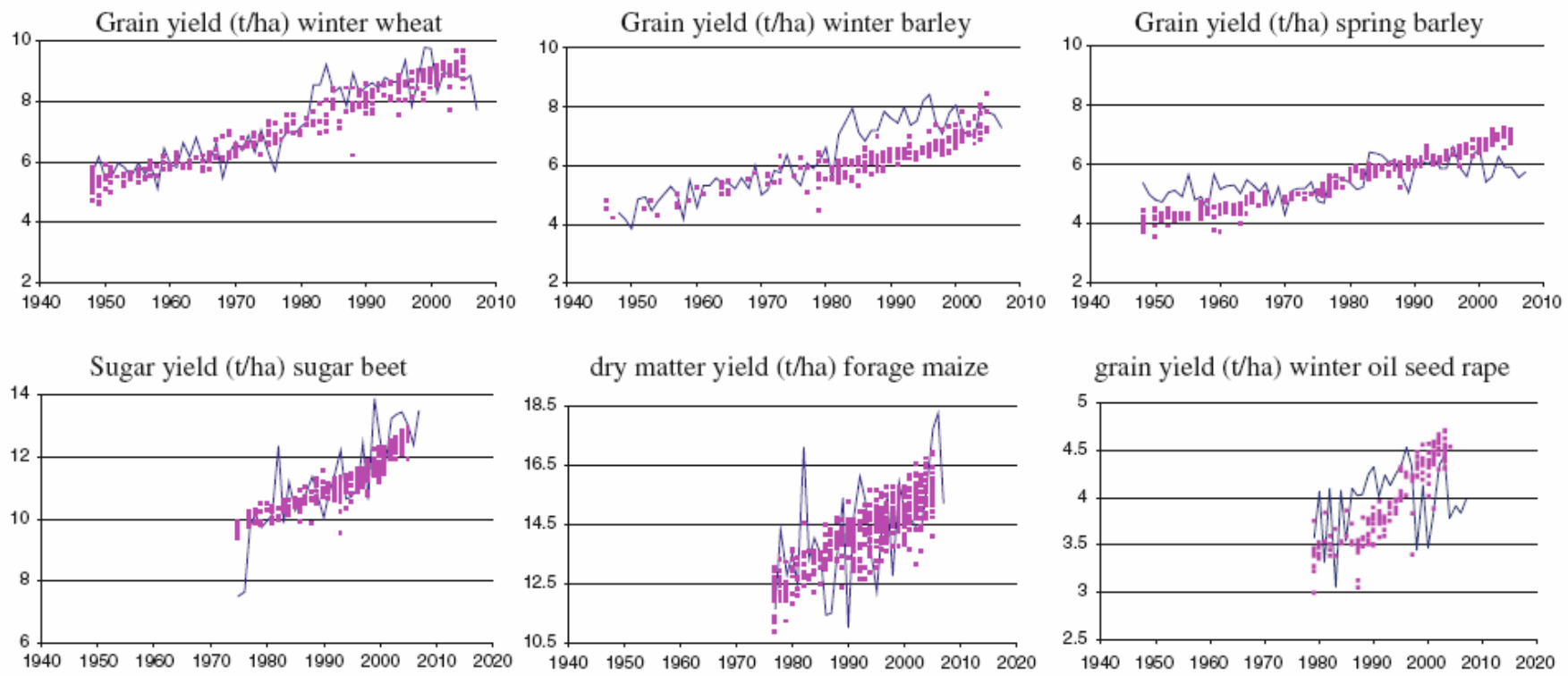

Fig. 1 Trends in variety and year effect for yield (t/ha) from 1948 to 2007. Ordinate and abscissa are on the same scale for all crops except oil seed rape. Variety and year means were estimated as described in "Materials and methods" section. Variety effects (*squares*) are plotted against the year in which the variety first entered the trial. Year means are plotted as a *line*

(Mackay et al., 2011)

—— year means

■ variety means

4. Example LMM

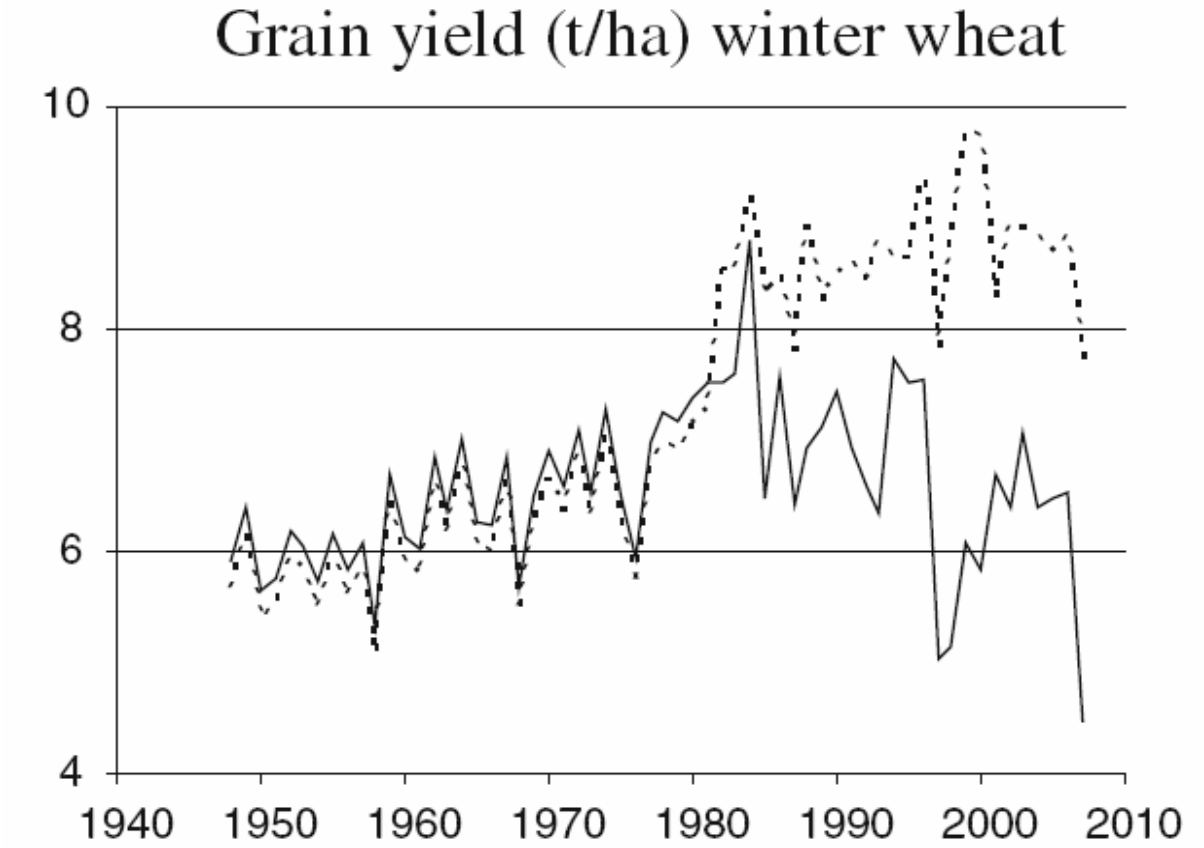

Fig. 3: Trends in year effects in UK variety trials (Mackay et al. 2011)

– – – – *Dotted line*: analysis with treated trials only (from 1982 onwards).

——— *Solid line*: analysis with untreated trials only (from 1982 onwards).

4. Example LMM

Genetic trend

$$G_i = \beta r_i + H_i \tag{15}$$

β = fixed regression coefficient for genetic trend

r_i = year of first trial for *i*-th variety

$H_i \sim N\left(0, \sigma_H^2\right)$

Non-genetic trend

$$Y_k = \gamma t_k + Z_k \tag{16}$$

γ = fixed regression coefficient for agronomic trend

t_k = calendar year

$Z_k \sim N\left(0, \sigma_Z^2\right)$

4. Example LMM

The fixed part of the model so far

$$\eta_{ik} = \mu + \beta r_i + \gamma t_k \qquad (17)$$

η_{ik} = expected response of the *i*-th genotype in the *k*-th year

Extension by environmental covariates

$$\eta_{ik} = \mu + \beta r_i + \gamma t_k + \delta_1 x_{kj1} + \delta_2 x_{kj2} + \delta_3 x_{kj3} + \delta_4 x_{kj4} \qquad (18)$$

δ_1 to δ_4 : regression slopes for the average mean temperature, precipitation, sunshine duration and drought index. (selected by cross validation) (Hadasch et al., 2020)

4. Example LMM

Table 2: Variance components (grain corn yield, Germany, 1992 to 2016)

Random effect	Null model	Time trend	Time trend + covariates
G	9.41	12.2	12.2
L	102.9	102.4	89.5
Y	129.5	58.9	29.9
L•Y	152.2	152.1	148.2
G•Y	11.9	4.5	4.5
G•L	3.9	3.9	3.9
G•L•Y	27.8	27.8	27.8

4. Example LMM

Table 3: Regression coefficient estimates (grain corn yield)

	Time trend		Time trend + covariates	
Fixed effect	Estimate	Standard error	Estimate	Standard error
μ	-2170	464	-2664	367
β	1.56	0.07	1.56	0.07
γ	0.417	0.24	-0.176	0.195
δ_1			-0.74	0.663
δ_2			0.067	0.037
δ_3			0.129	0.037
δ_4			2.065	0.511

4. Example LMM

Table 4: Coefficients of determination Ω_β

Trait	θ^{AMV}		θ^{ASV}	
	Time	Time + cov.	Time	Time + cov.
Yield	0.17296	0.27745	0.16935	0.27231
Height	0.09883	0.16525	0.09696	0.16263
Lodging (%)	0.00378	0.01794	0.00375	0.01721
Thousand kernel weight	0.01730	0.02967	0.01682	0.02889
Dry matter	0.01556	0.31758	0.01516	0.31413

5. Extension to generalized linear mixed models (GLMM)

$$\eta = X\beta + Zu \tag{19}$$

$$E(y \mid \eta) = \mu = g^{-1}(\eta), \tag{20}$$

where $g(.)$ is a link function.

Problem:

On which scale to assess total variance?

5. Extension to generalized linear mixed models (GLMM)

Assessing total variance on linear predictor scale

$\Rightarrow$ generally add a random unit effect f with zero mean and $\mathrm{var}(f) = R_f$:

$$\eta = X\beta + Zu + f \qquad (21)$$

$\Rightarrow$ add auxiliary random residual vector $h^T = (h_1, h_2, \ldots)$:

$$\tilde{\eta} = \eta + h$$

with conditional variance $\mathrm{var}\left(g^{-1}(\tilde{\eta}) \mid \eta\right)$

(Nakagawa and Schielzeth, 2013)

5. Extension to generalized linear mixed models (GLMM)

Now ask:

Which variance-covariance $R_h = \mathrm{var}(h)$ leads to $\mathrm{var}\left(g^{-1}(\tilde{\eta}) \mid \eta\right) \approx \mathrm{var}(y \mid \mu)$?

Total variance-covariance matrix:

$$\tilde{V} = ZGZ^T + \tilde{R} \tag{22}$$

with

$$\tilde{R} = R_f + R_h \tag{23}$$

(Foulley et al., 1987)

5. Extension to generalized linear mixed models (GLMM)

Special cases, where $\mathrm{var}\left(g^{-1}(\tilde{\eta})\,|\,\eta\right) = \mathrm{var}(y\,|\,\mu)$ exactly:

(1) Binomial distribution, logit link

$\Rightarrow$ logistic distribution may be assumed for h_i with $\mathrm{var}(h_i) = \pi^2/3$

(2) Binomial distribution, probit link

$\Rightarrow$ standard normal distribution for h_i with $\mathrm{var}(h_i) = 1$

(Keen and Engel, 1997)

5. Extension to generalized linear mixed models (GLMM)

Other distributions and links:

$\Rightarrow$ take recourse to an approximation based on a Taylor series expansion

Assume:

$$\mathrm{var}(y \mid \mu) = A_{\mu}^{1/2} R A_{\mu}^{1/2} \,, \tag{24}$$

where

A_{μ} = diagonal matrix with evaluations of the variance function at mean μ

R = known or unknown matrix

(Wolfinger and O'Connell, 1993)

5. Extension to generalized linear mixed models (GLMM)

Special cases:

(1) LMM $\Rightarrow$ identity link and $A_\mu = I$

(2) GLMMs with conditional error distribution in exponential family $\Rightarrow$ $R = I_n$

It may be assumed that

$$\mathrm{var}(h) = R_h = W_\mu^{1/2} R W_\mu^{1/2} \,, \qquad (25)$$

where W_μ is a diagonal matrix with functions of the mean μ on the diagonal.

5. Extension to generalized linear mixed models (GLMM)

Expanding $g^{-1}(\tilde{\eta})$ in a Taylor series about the mean η of the linear predictor, we find that to first order

$$\mathrm{var}\left(g^{-1}(\tilde{\eta})\,|\,\eta\right) \approx D_{\eta^*} W_\mu^{1/2} R W_\mu^{1/2} D_{\eta^*} \quad , \tag{26}$$

where $D_{\eta^*} = diag\left[\partial g^{-1}(\tilde{\eta})/\partial\tilde{\eta}\right]_{\tilde{\eta}=\eta}$.

Comparing coefficients between (21) and (23) yields $A_\mu = D_{\tilde{\eta}} W_\mu D_{\tilde{\eta}}$ and hence $W_\mu = D_{\tilde{\eta}}^{-1} A_\mu D_{\tilde{\eta}}^{-1}$.

(Foulley et al., 1987; Wolfinger and O'Connell, 1993; Bennewitz et al., 2014)

5. Extension to generalized linear mixed models (GLMM)

Examples

(1) Overdispersed Poisson data with log-link:

$$\mathrm{var}(y_i \mid \mu_i) = \phi\mu_i \Rightarrow \mathrm{var}(h_i) \approx \phi\mu_i^{-1} .$$

(Foulley et al., 1987)

(2) Overdispersed binomial data with probit link:

$$\mathrm{var}(y_i \mid \mu_i) = \frac{\phi\mu_i(1-\mu_i)}{m_i} \Rightarrow \mathrm{var}(h_i) \approx \phi\frac{\mu_i(1-\mu_i)}{[\varphi(\eta_i)]^2 m_i} ,$$

where m_i = binomial sample size of the *i*-th observation and

$\varphi(.)$ = standard normal probability density.

(Bennewitz et al., 2014)

6. Summary

Coefficient of determination for LMM can be defined based on pairwise differences and semivariances.

There are several proposals in the market. Not all of them are easy to interpret and communicate.

I think that pairwise differences are very easy to communicate to users and they make a lot of sense.

References

Bennewitz, J., Böglein, S., Stratz, P., Rodehutscord, M., Piepho, H.P., Kjaer, W., Bessei, W. (2014): Genetic parameters for feather pecking and aggressive behaviour in a large F2-cross of laying hens using generalized linear mixed models. *Poultry Science* **93**, 810-817.

Feldmann, M.J., Piepho, H.P., Bridges, W.C., Knapp, S.J. (2021): Accurate estimation of marker-associated genetic variance and heritability in complex trait analyses. *bioRxiv*

Foulley, J. L., Gianola, D. and Im, S. (1987): Genetic evaluation of traits distributed as Poisson-binomial with reference to reproduction characters. *Theor. Appl. Genet.* **73**, 870–877.

Hadasch, S., Laidig, F., Macholdt, J., Bönecke, E., Piepho, H.P. (2020): Trends in the mean performance and stability of winter wheat and winter barley yields using a long-term series of variety trials. *Field Crops Research* **251**, 107792.

John, J.A., Williams, E.R. (1995): *Cyclic and computer generated designs* (2nd ed.). London, UK: Chapman and Hall.

Keen, A. and Engel, B. (1997): Analysis of a mixed model for ordinal data by iterative re-weighted least squares. *Statistica Neerlandica* **51**, 129–144.

Mackay, I.J., Horwell, A., Garner, J., White, J., McKee, J., Philpott, H. (2011): Reanalysis of the historical series of UK variety trials to quantify the contributions of genetic and environmental factors to trends and variability in yield over time. *TAG* **122**, 225–238.

Nakagawa, S., Schielzeth, H. (2013): A general and simple method for obtaining R^2 from generalized linear mixed models. *Methods in Ecology and Evolution* **4**, 133–142.

Piepho, H.P., Möhring, J. (2007): Computing heritability and selection response from unbalanced plant breeding trials. *Genetics* **177**, 1881-1888.

Piepho, H.P. (2019): A coefficient of determination (R^2) for generalized linear mixed models. *Biometrical Journal* **61**, 860-872.

Webster, R., Oliver, M.E. (2007): Geostatistics for environmental scientists. Wiley, New York.

Wolfinger, R. and O'Connell, M. (1993): Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* **48**, 233–243.