Assessing disease resistance in chickpeas through the bivariate analysis of Gaussian and binomial traits.

Aidan McGarty[1]
Brian Cullis[1], Ahsan Asif[2] and Kristy Hobson[2]

November 29, 2023

Mixed Models and Experiment Design Lab (MMaEDL)[1]
National Institute for Applied Statistics Research Australia
University of Wollongong
`amcgarty@uow.edu.au`

Chickpea Breeding Australia[2]
NSW Department of Primary Industries | Agriculture

Overview

- **Motivation**
  - Context
  - Complexities
- **The statistics of the problem**
  - Choice of model
- **Results**
  - Parametric bootstrapping

Motivation

Motivation

- Investigate disease (Phytophthora root rot - PRR) resistance in Australian chickpeas

Motivation

- Investigate disease (Phytophthora root rot - PRR) resistance in Australian chickpeas
- Generally done through field experiments as these most similarly match conditions Australian chickpea growers experience

Motivation

- Investigate disease (Phytophthora root rot - PRR) resistance in Australian chickpeas
- Generally done through field experiments as these most similarly match conditions Australian chickpea growers experience
- However there are constraints to manage:

Motivation

- Investigate disease (Phytophthora root rot - PRR) resistance in Australian chickpeas
- Generally done through field experiments as these most similarly match conditions Australian chickpea growers experience
- However there are constraints to manage:
    - Difficult to control disease pressure in field experiments

Motivation

- Investigate disease (Phytophthora root rot - PRR) resistance in Australian chickpeas
- Generally done through field experiments as these most similarly match conditions Australian chickpea growers experience
- However there are constraints to manage:
    - Difficult to control disease pressure in field experiments
    - Difficult to directly measure/quantify disease presence

Motivation

- Investigate disease (Phytophthora root rot - PRR) resistance in Australian chickpeas
- Generally done through field experiments as these most similarly match conditions Australian chickpea growers experience
- However there are constraints to manage:
  - Difficult to control disease pressure in field experiments
  - Difficult to directly measure/quantify disease presence
  - Field experiments are expensive, labour intensive and time consuming

Motivation

- Investigate disease (Phytophthora root rot - PRR) resistance in Australian chickpeas
- Generally done through field experiments as these most similarly match conditions Australian chickpea growers experience
- However there are constraints to manage:
    - Difficult to control disease pressure in field experiments
    - Difficult to directly measure/quantify disease presence
    - Field experiments are expensive, labour intensive and time consuming
- These factors lead to a complex statistical analysis

Motivation

Scenario

Scenario

- This investigation involves data from two experiments - one field experiment and one hydroponic experiment

Scenario

- This investigation involves data from two experiments - one field experiment and one hydroponic experiment
- Both experiments assess a similar set of genetic material (varieties) under artificially induced/amplified PRR pressure

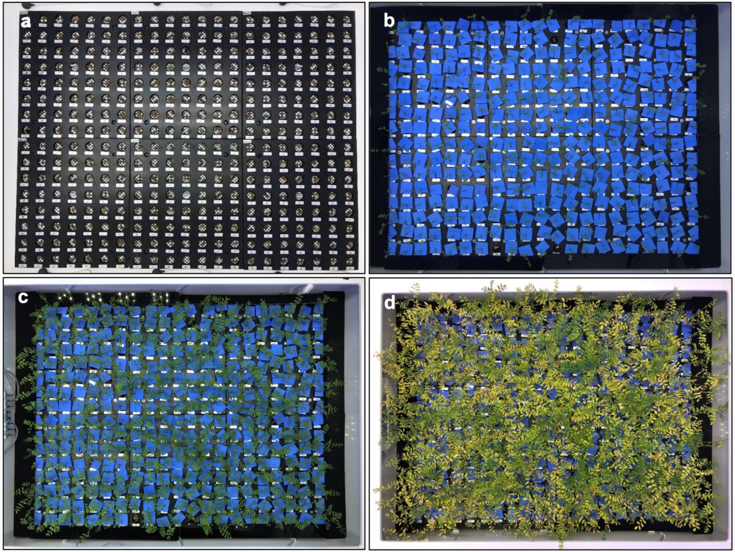
Scenario

- This investigation involves data from two experiments - one field experiment and one hydroponic experiment
- Both experiments assess a similar set of genetic material (varieties) under artificially induced/amplified PRR pressure
- Researcher would like to assess the agreement between PRR resistance assessment in the hydroponic and field experiments

Scenario

- This investigation involves data from two experiments - one field experiment and one hydroponic experiment
- Both experiments assess a similar set of genetic material (varieties) under artificially induced/amplified PRR pressure
- Researcher would like to assess the agreement between PRR resistance assessment in the hydroponic and field experiments
- Hydroponics experiments provide high throughput low cost alternatives to field experiments

Scenario - Hydroponic



Scenario - Hydroponic

Scenario - Hydroponic

- In the hydroponic experiment counts of total and chlorotic leaves per plant are recorded

Scenario - Hydroponic

- In the hydroponic experiment counts of total and chlorotic leaves per plant are recorded
- Along with terms associated with the plot structure; Tank, Racks, Rows and Columns

Scenario - Hydroponic

- In the hydroponic experiment counts of total and chlorotic leaves per plant are recorded
- Along with terms associated with the plot structure; Tank, Racks, Rows and Columns

Tank	Rack	HRange	HRow	Variety	TotalLeaf	DisLeaf
1	1	1	1	A	6	2
1	1	2	1	B	5	3
1	1	3	1	C		
1	1	4	1	D	11	6
1	1	5	1	F	13	9
1	1	6	1	A	15	7

Scenario - Field

Scenario - Field

Scenario - Field

- In the field experiment dead and chlorotic plants along with total plants at risk per plot are recorded at three different time points

Scenario - Field

- In the field experiment dead and chlorotic plants along with total plants at risk per plot are recorded at three different time points
- This in itself introduces complexities however following the methods of Bartlett (1978), expected lifetimes have been calculated for each plot (Amalraj et al., 2019)

Scenario - Field

- In the field experiment dead and chlorotic plants along with total plants at risk per plot are recorded at three different time points
- This in itself introduces complexities however following the methods of Bartlett (1978), expected lifetimes have been calculated for each plot (Amalraj et al., 2019)
- Terms associated with the field experiment plot structure are also recorded; Rows and Columns

Scenario - Field

- In the field experiment dead and chlorotic plants along with total plants at risk per plot are recorded at three different time points
- This in itself introduces complexities however following the methods of Bartlett (1978), expected lifetimes have been calculated for each plot (Amalraj et al., 2019)
- Terms associated with the field experiment plot structure are also recorded; Rows and Columns

FCol	FRow	Variety	Risk1	Dead1	Risk2	Dead2	Risk3	Dead3	ExpectedLife
1	1	A	16	0	16	12	4	15	1.57
1	2	B	16	0	16	0	16	0	9.39
1	3	C	16	0	16	0	16	2	4.01
1	4	D	10	0	10	7	3	9	1.75
1	5	E	19	0	19	2	17	7	2.98
1	6	A	16	0	16	4	12	7	2.87

Statistics of the Problem

Statistics of the Problem

- We would like a statistical model which:

Statistics of the Problem

- We would like a statistical model which:
    - Reflects the plot structure of each experiment

Statistics of the Problem

- We would like a statistical model which:
    - Reflects the plot structure of each experiment
    - Accurately captures the underlying distribution of each measured trait

Statistics of the Problem

- We would like a statistical model which:
    - Reflects the plot structure of each experiment
    - Accurately captures the underlying distribution of each measured trait
    - Provides a reliable estimate of the genetic correlation between experiments

Statistics of the Problem

- We would like a statistical model which:
    - Reflects the plot structure of each experiment
    - Accurately captures the underlying distribution of each measured trait
    - Provides a reliable estimate of the genetic correlation between experiments
- We choose a bivariate generalised linear mixed model (GLMM) using ASReml-R

Statistics of the Problem

Statistics of the Problem

- ASReml-R implements penalised quasilikelihood (PQL) estimation (Breslow & Clayton, 1993) which allows us to fit such a model following some manipulation of the dataframe

Statistics of the Problem

- ASReml-R implements penalised quasilikelihood (PQL) estimation (Breslow & Clayton, 1993) which allows us to fit such a model following some manipulation of the dataframe
- Currently the only method to fit multivariate GLMMs in ASReml-R is through specifying equal design matrices among the traits (experiments)

Statistics of the Problem

- ASReml-R implements penalised quasilikelihood (PQL) estimation (Breslow & Clayton, 1993) which allows us to fit such a model following some manipulation of the dataframe
- Currently the only method to fit multivariate GLMMs in ASReml-R is through specifying equal design matrices among the traits (experiments)
- This creates complexities in this scenario, as the two experiments have different sets of terms associated with their plot structures

Statistics of the Problem

- ASReml-R implements penalised quasilikelihood (PQL) estimation (Breslow & Clayton, 1993) which allows us to fit such a model following some manipulation of the dataframe
- Currently the only method to fit multivariate GLMMs in ASReml-R is through specifying equal design matrices among the traits (experiments)
- This creates complexities in this scenario, as the two experiments have different sets of terms associated with their plot structures
- To overcome this the two data frames are merged by variety in a non-unique way

Statistics of the Problem

Statistics of the Problem

WPlot	Variety	FCol	FRow	Tank	Rack	HRange	HRow	DL*	TL*	EL*
1	A	1	1	1	1	1	1	2	6	1.57
2	B	1	2	1	1	2	1	3	5	9.39
3	C	1	3	1	1	3	1			4.01
4	D	1	4	1	1	4	1	6	11	1.75
5	E	1	5							2.98
6	F			1	1	5	1	9	13	
7	A	1	6	1	1	6	1	7	15	2.87

*DL = Number of chlorotic leaves (hydroponic), *TL = Number of total leaves (hydroponic) and *EL = Expected lifetime (field)

Statistics of the Problem

WPlot	Variety	FCol	FRow	Tank	Rack	HRange	HRow	DL*	TL*	EL*
1	A	1	1	1	1	1	1	2	6	1.57
2	B	1	2	1	1	2	1	3	5	9.39
3	C	1	3	1	1	3	1			4.01
4	D	1	4	1	1	4	1	6	11	1.75
5	E	1	5							2.98
6	F			1	1	5	1	9	13	
7	A	1	6	1	1	6	1	7	15	2.87

*DL = Number of chlorotic leaves (hydroponic), *TL = Number of total leaves (hydroponic) and *EL = Expected lifetime (field)

- WPlot arbitrarily joins field and hydroponic observations by common varieties, however as evident in WPlot 5 and 6 not all field and hydroponic observations can be matched

Model

Model

- As previously discussed, we have data from two experiments $\mathbf{y} = (\mathbf{y}_H^\intercal, \mathbf{y}_F^\intercal)^\intercal$:

Model

- As previously discussed, we have data from two experiments $\mathbf{y} = (\mathbf{y}_H^\intercal, \mathbf{y}_F^\intercal)^\intercal$:
    - $\mathbf{y}_H = (y_{H_1}, \ldots, y_{H_{n_H}})^\intercal$ from the hydroponic experiment which we assume follows a binomial distribution

Model

- As previously discussed, we have data from two experiments $\mathbf{y} = (\mathbf{y}_H^\intercal, \mathbf{y}_F^\intercal)^\intercal$:
  - $\mathbf{y}_H = (y_{H_1}, \ldots, y_{H_{n_H}})^\intercal$ from the hydroponic experiment which we assume follows a binomial distribution
  - $\mathbf{y}_F = (y_{F_1}, \ldots, y_{F_{n_F}})^\intercal$ from the field experiment which we assume follows a normal distribution

Model

- As previously discussed, we have data from two experiments $\mathbf{y} = (\mathbf{y}_H^\intercal, \mathbf{y}_F^\intercal)^\intercal$:
  - $\mathbf{y}_H = (y_{H_1}, \ldots, y_{H_{n_H}})^\intercal$ from the hydroponic experiment which we assume follows a binomial distribution
  - $\mathbf{y}_F = (y_{F_1}, \ldots, y_{F_{n_F}})^\intercal$ from the field experiment which we assume follows a normal distribution
- We would also like to include random effects $\mathbf{u} = (\mathbf{u}_g^\intercal, \mathbf{u}_{p_H}^\intercal, \mathbf{u}_{p_F}^\intercal)^\intercal$ which are either shared genetic effects $\mathbf{u}_g$ or non-shared peripheral effects $\mathbf{u}_{p_H}$ and $\mathbf{u}_{p_F}$, for the hydroponic and field experiments respectively

Model

- As previously discussed, we have data from two experiments $\mathbf{y} = (\mathbf{y}_H^\intercal, \mathbf{y}_F^\intercal)^\intercal$:
    - $\mathbf{y}_H = (y_{H_1}, \ldots, y_{H_{n_H}})^\intercal$ from the hydroponic experiment which we assume follows a binomial distribution
    - $\mathbf{y}_F = (y_{F_1}, \ldots, y_{F_{n_F}})^\intercal$ from the field experiment which we assume follows a normal distribution
- We would also like to include random effects $\mathbf{u} = (\mathbf{u}_g^\intercal, \mathbf{u}_{p_H}^\intercal, \mathbf{u}_{p_F}^\intercal)^\intercal$ which are either shared genetic effects $\mathbf{u}_g$ or non-shared peripheral effects $\mathbf{u}_{p_H}$ and $\mathbf{u}_{p_F}$, for the hydroponic and field experiments respectively
- We can then write the distributions $\mathbf{y}_H|\mathbf{u}$ and $\mathbf{y}_F|\mathbf{u}$ conditional on the random effects $\mathbf{u}$ with associated probability density functions (PDF) $f_{Y_H|U}$ and $f_{Y_F|U}$

Model - Hydroponic Conditional Distribution

$$E(y_{H_i}|\mathbf{u}) = \mu_{H_i}$$

$$g(\mu_{H_i}) = \eta_{H_i} = \mathbf{x}_{H_i}^\top \boldsymbol{\tau} + \mathbf{z}_{H_i}^\top \mathbf{u}$$

$$V(y_{H_i}|\mathbf{u}) = \phi_H \mu_{H_i}(1 - \frac{\mu_{H_i}}{n_{TL_i}})$$

- for $i = 1, \ldots, n_H$ where $g()$ is the logit link function
- $\mathbf{x}_{H_i}$ and $\mathbf{z}_{H_i}$ are indicator/covariate vectors of length c_X and c_Z relating to the fixed ($\boldsymbol{\tau}$) and random ($\mathbf{u}$) effects for the ith observation in the hydroponic experiment
- ϕ_H represents the dispersion parameter and n_{TL_i} are the binomial totals for the ith observation in the hydroponic experiment

Model - Field Conditional Distribution

$$E(y_{F_i}|\mathbf{u}) = \eta_{F_i} = \mathbf{x}_{F_i}^{\top}\boldsymbol{\tau} + \mathbf{z}_{F_i}^{\top}\mathbf{u}$$

$$V(y_{F_i}|\mathbf{u}) = \sigma_F^2$$

- for $i = 1, \ldots, n_F$
- $\mathbf{x}_{F_i}$ and $\mathbf{z}_{F_i}$ are indicator/covariate vectors of length c_X and c_Z relating to the fixed ($\boldsymbol{\tau}$) and random ($\mathbf{u}$) effects for the ith observation in the field experiment
- σ_F^2 is the field residual variance

Random Effects

Random Effects

- We assume the random effects follow a multivariate normal distribution such that:

Random Effects

- We assume the random effects follow a multivariate normal distribution such that:

$$\begin{bmatrix} \mathbf{u}_g \\ \mathbf{u}_{p_H} \\ \mathbf{u}_{p_F} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G}_g & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{p_H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{G}_{p_H} \end{bmatrix} \right)$$

Random Effects

- We assume the random effects follow a multivariate normal distribution such that:

$$\begin{bmatrix} \mathbf{u}_g \\ \mathbf{u}_{p_H} \\ \mathbf{u}_{p_F} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G}_g & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{p_H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{G}_{p_H} \end{bmatrix} \right)$$

- Or more generally $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$ with associated PDF f_U

Random Effects

- We assume the random effects follow a multivariate normal distribution such that:

$$\begin{bmatrix} \mathbf{u}_g \\ \mathbf{u}_{p_H} \\ \mathbf{u}_{p_F} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G}_g & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{p_H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{G}_{p_H} \end{bmatrix} \right)$$

- Or more generally $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$ with associated PDF f_U
- For convenience we allow $\boldsymbol{\kappa}$ to contain all parameters in $\mathbf{G}_g$, $\mathbf{G}_{p_H}$ and $\mathbf{G}_{p_F}$ along with ϕ_H and σ_F^2

Likelihood

Likelihood

- Our interest then lies in the log-likelihood, which following Collins (2008) can be written as:

Likelihood

- Our interest then lies in the log-likelihood, which following Collins (2008) can be written as:

$$\begin{aligned}\ell(\boldsymbol{\tau}, \boldsymbol{\kappa}; \mathbf{y}) &= \ell(\boldsymbol{\tau}, \boldsymbol{\kappa}; \mathbf{y}_H) + \ell(\boldsymbol{\tau}, \boldsymbol{\kappa}; \mathbf{y}_F) \\ &= \log\left(\int f_{Y_H|U} f_U d\mathbf{u}\right) + \log\left(\int f_{Y_F|U} f_U d\mathbf{u}\right)\end{aligned}$$

Likelihood

- Our interest then lies in the log-likelihood, which following Collins (2008) can be written as:

$$\begin{aligned}\ell(\boldsymbol{\tau}, \boldsymbol{\kappa}; \mathbf{y}) &= \ell(\boldsymbol{\tau}, \boldsymbol{\kappa}; \mathbf{y}_H) + \ell(\boldsymbol{\tau}, \boldsymbol{\kappa}; \mathbf{y}_F) \\ &= \log\left(\int f_{Y_H|U} f_U d\mathbf{u}\right) + \log\left(\int f_{Y_F|U} f_U d\mathbf{u}\right)\end{aligned}$$

- The first term of which is not analytically tractable and hence requires an alternative to classical likelihood inference - therefore PQL is used to approximate the likelihood (Collins, 2008)

Model - Variance Structure of Random Effects

Model - Variance Structure of Random Effects

- The key component of this model is the variance model $\mathbf{G}_g$ for the set of genetic effects $\mathbf{u}_g$

Model - Variance Structure of Random Effects

- The key component of this model is the variance model $\mathbf{G}_g$ for the set of genetic effects $\mathbf{u}_g$
- The total genetic effect are partitioned such that $\mathbf{u}_g = \mathbf{u}_a + \mathbf{u}_e$ into an additive ($\mathbf{u}_a$) and non-additive component ($\mathbf{u}_e$)

Model - Variance Structure of Random Effects

- The key component of this model is the variance model $\mathbf{G}_g$ for the set of genetic effects $\mathbf{u}_g$
- The total genetic effect are partitioned such that $\mathbf{u}_g = \mathbf{u}_a + \mathbf{u}_e$ into an additive ($\mathbf{u}_a$) and non-additive component ($\mathbf{u}_e$)

$$\begin{bmatrix} \mathbf{u}_a \\ \mathbf{u}_e \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G}_{g_a} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{g_e} \end{bmatrix} \right)$$

Model - Variance Structure of Random Effects

- The key component of this model is the variance model $\mathbf{G}_g$ for the set of genetic effects $\mathbf{u}_g$
- The total genetic effect are partitioned such that $\mathbf{u}_g = \mathbf{u}_a + \mathbf{u}_e$ into an additive ($\mathbf{u}_a$) and non-additive component ($\mathbf{u}_e$)

$$\begin{bmatrix} \mathbf{u}_a \\ \mathbf{u}_e \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G}_{g_a} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{g_e} \end{bmatrix} \right)$$

- Here $\mathbf{G}_{g_a} = \boldsymbol{\Sigma}_a \otimes \mathbf{K}$ and $\mathbf{G}_{g_e} = \boldsymbol{\Sigma}_e \otimes \mathbf{I}_{c_g}$ where $\mathbf{K}$ is a $(c_g \times c_g)$ known genomic relationship matrix (GRM) formed via marker scores

Model - Variance Structure of Random Effects

- The key component of this model is the variance model $\mathbf{G}_g$ for the set of genetic effects $\mathbf{u}_g$
- The total genetic effect are partitioned such that $\mathbf{u}_g = \mathbf{u}_a + \mathbf{u}_e$ into an additive ($\mathbf{u}_a$) and non-additive component ($\mathbf{u}_e$)

$$\begin{bmatrix} \mathbf{u}_a \\ \mathbf{u}_e \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G}_{g_a} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{g_e} \end{bmatrix} \right)$$

- Here $\mathbf{G}_{g_a} = \mathbf{\Sigma}_a \otimes \mathbf{K}$ and $\mathbf{G}_{g_e} = \mathbf{\Sigma}_e \otimes \mathbf{I}_{c_g}$ where $\mathbf{K}$ is a $(c_g \times c_g)$ known genomic relationship matrix (GRM) formed via marker scores
- Where $\mathbf{\Sigma}_a = \begin{bmatrix} \sigma_{a_H} & \sigma_{a_{HF}} \\ \sigma_{a_{HF}} & \sigma_{a_F} \end{bmatrix}$ and $\mathbf{\Sigma}_e = \begin{bmatrix} \sigma_{e_H} & \sigma_{e_{HF}} \\ \sigma_{e_{HF}} & \sigma_{e_F} \end{bmatrix}$ is known as an unstructured parameterisation

Model - Variance Structure of Random Effects

- The key component of this model is the variance model $\mathbf{G}_g$ for the set of genetic effects $\mathbf{u}_g$
- The total genetic effect are partitioned such that $\mathbf{u}_g = \mathbf{u}_a + \mathbf{u}_e$ into an additive ($\mathbf{u}_a$) and non-additive component ($\mathbf{u}_e$)

$$\begin{bmatrix} \mathbf{u}_a \\ \mathbf{u}_e \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G}_{g_a} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{g_e} \end{bmatrix} \right)$$

- Here $\mathbf{G}_{g_a} = \boldsymbol{\Sigma}_a \otimes \mathbf{K}$ and $\mathbf{G}_{g_e} = \boldsymbol{\Sigma}_e \otimes \mathbf{I}_{c_g}$ where $\mathbf{K}$ is a $(c_g \times c_g)$ known genomic relationship matrix (GRM) formed via marker scores
- Where $\boldsymbol{\Sigma}_a = \begin{bmatrix} \sigma_{a_H} & \sigma_{a_{HF}} \\ \sigma_{a_{HF}} & \sigma_{a_F} \end{bmatrix}$ and $\boldsymbol{\Sigma}_e = \begin{bmatrix} \sigma_{e_H} & \sigma_{e_{HF}} \\ \sigma_{e_{HF}} & \sigma_{e_F} \end{bmatrix}$ is known as an unstructured parameterisation
- In the case of the non-genetic effects, we specify $\mathbf{G}_{p_H}$ and $\mathbf{G}_{p_F}$ as block diagonal matrices

The Model - Genetic Effects

The Model - Genetic Effects

- Computationally, estimating the parameters in the unstructured formulation may lead to estimates outside the parameter space (Meyer, 2023)

The Model - Genetic Effects

- Computationally, estimating the parameters in the unstructured formulation may lead to estimates outside the parameter space (Meyer, 2023)
- We use a slightly different parameterisation of the genetic variance matrices:

The Model - Genetic Effects

- Computationally, estimating the parameters in the unstructured formulation may lead to estimates outside the parameter space (Meyer, 2023)
- We use a slightly different parameterisation of the genetic variance matrices:

$$\boldsymbol{\Sigma}_a = \begin{bmatrix} \lambda_{a_H}^2 & \lambda_{a_H}\lambda_{a_F} \\ \lambda_{a_H}\lambda_{a_F} & \lambda_{a_F}^2 + \psi_{a_F} \end{bmatrix} \text{ and } \boldsymbol{\Sigma}_e = \begin{bmatrix} \lambda_{e_H}^2 & \lambda_{e_H}\lambda_{e_F} \\ \lambda_{e_H}\lambda_{e_F} & \lambda_{e_F}^2 + \psi_{e_F} \end{bmatrix}$$

The Model - Genetic Effects

- Computationally, estimating the parameters in the unstructured formulation may lead to estimates outside the parameter space (Meyer, 2023)
- We use a slightly different parameterisation of the genetic variance matrices:
  $\boldsymbol{\Sigma}_a = \begin{bmatrix} \lambda_{a_H}^2 & \lambda_{a_H}\lambda_{a_F} \\ \lambda_{a_H}\lambda_{a_F} & \lambda_{a_F}^2 + \psi_{a_F} \end{bmatrix}$ and $\boldsymbol{\Sigma}_e = \begin{bmatrix} \lambda_{e_H}^2 & \lambda_{e_H}\lambda_{e_F} \\ \lambda_{e_H}\lambda_{e_F} & \lambda_{e_F}^2 + \psi_{e_F} \end{bmatrix}$
- This parameterisation is known as "reduced rank + diag" and is convenient as estimates of the λ_{i_j}, $i = a, e$, $j = H, F$ cannot go outside the parameter space unlike the "σ" parameters in the unstructured parameterisation

Model - Recap

Model - Recap

- Due to the non-Gaussian nature of the response recorded in the hydroponic experiment we use PQL, an approximate likelihood method as the likelihood analytically intractable

Model - Recap

- Due to the non-Gaussian nature of the response recorded in the hydroponic experiment we use PQL, an approximate likelihood method as the likelihood analytically intractable
- For the genetic effects, which are our main interest, we specify a variance structure which allows for covariance between experiments

Model - ASReml

```r
asreml(fixed = cbind(DL,EL) ~ trait + trait:VarietyDrop,
       random =~ rr(trait):vm(Variety, K) + at(trait,"EL"):vm(Variety, K) +
         rr(trait):ide(Variety) + at(trait,"EL"):ide(Variety) +
         at(trait, "DL"):Tank + at(trait, "DL"):Tank:Rack +
         at(trait, "DL"):Tank:Rack:HRow + at(trait, "DL"):Tank:Rack:HRange +
         at(trait, "EL"):FRow + at(trait, "EL"):FCol,
       residual = ~id(WPlot):diag(trait),
       data = df,
       family = list(asr_binomial(total = "TL"), asr_gaussian()),
       na.action = na.method(x="include",y="include"))
```

Results

Results

- PQL estimate of total genetic correlation between the two experiments of -0.476

Results

- PQL estimate of total genetic correlation between the two experiments of -0.476
- Negative correlations are expected, as for the field experiment higher expected lifetime values indicate higher disease resistance, whereas for the hydroponic experiment lower leaf chlorosis incidence values indicate higher disease resistance

Results

- PQL estimate of total genetic correlation between the two experiments of -0.476
- Negative correlations are expected, as for the field experiment higher expected lifetime values indicate higher disease resistance, whereas for the hydroponic experiment lower leaf chlorosis incidence values indicate higher disease resistance
- The researcher would also like a confidence interval on this estimate

Results

- PQL estimate of total genetic correlation between the two experiments of -0.476
- Negative correlations are expected, as for the field experiment higher expected lifetime values indicate higher disease resistance, whereas for the hydroponic experiment lower leaf chlorosis incidence values indicate higher disease resistance
- The researcher would also like a confidence interval on this estimate
- To do so while also assessing potential bias in the PQL estimates, a known issue impacting parameter estimation (Breslow & Lin, 1995) we implement parametric bootstrapping

Parametric Bootstrapping - Applied

Parametric Bootstrapping - Applied

- A total of 2000 bootstrap simulations were generated, of these 726 were omitted due to non-convergence

Parametric Bootstrapping - Applied

- A total of 2000 bootstrap simulations were generated, of these 726 were omitted due to non-convergence
- This was based on a criteria for parameter convergence;

Parametric Bootstrapping - Applied

- A total of 2000 bootstrap simulations were generated, of these 726 were omitted due to non-convergence
- This was based on a criteria for parameter convergence;

$$\Delta_{\text{norm}} = \sqrt{\frac{\sum_i (\hat{\theta}_i^{(t)} - \hat{\theta}_i^{(t-1)})^2}{\sum_i \hat{\theta}_i^{(t)2}}}$$

Parametric Bootstrapping - Applied

- A total of 2000 bootstrap simulations were generated, of these 726 were omitted due to non-convergence
- This was based on a criteria for parameter convergence;

$$\Delta_{\text{norm}} = \sqrt{\frac{\sum_i (\hat{\theta}_i^{(t)} - \hat{\theta}_i^{(t-1)})^2}{\sum_i \hat{\theta}_i^{(t)2}}}$$

- $i = 1, \ldots, p$ where p is the number of parameters, $\hat{\theta}_i^{(j)}$ is the estimate of the ith parameter at the jth iteration with t being the final iteration

Parametric Bootstrapping - Applied

- A total of 2000 bootstrap simulations were generated, of these 726 were omitted due to non-convergence
- This was based on a criteria for parameter convergence;

$$\Delta_{\text{norm}} = \sqrt{\frac{\Sigma_i(\hat{\theta}_i^{(t)} - \hat{\theta}_i^{(t-1)})^2}{\Sigma_i \hat{\theta}_i^{(t)2}}}$$

- $i = 1, \ldots, p$ where p is the number of parameters, $\hat{\theta}_i^{(j)}$ is the estimate of the ith parameter at the jth iteration with t being the final iteration
- After 50 iterations simulations where $\Delta_{\text{norm}} > 0.001$ were excluded

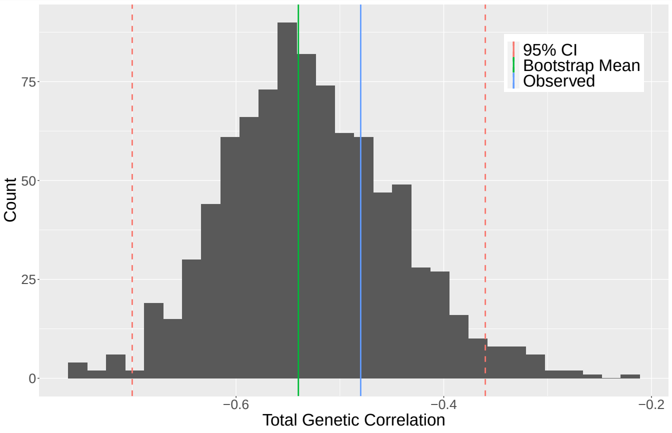
Parametric Bootstrapping - Applied

Parametric Bootstrapping - Applied

- The remaining 1274 simulations resulting in an average total genetic correlation of -0.54 with an associated 95% confidence interval (-0.70,-0.36)

Parametric Bootstrapping - Applied

- The remaining 1274 simulations resulting in an average total genetic correlation of -0.54 with an associated 95% confidence interval (-0.70,-0.36)



Parametric Bootstrapping - Sampling Distribution

Parameter	Obs	BootMean
σ^2_{Tank}	0.082	0.077
σ^2_{Rack}	0.022	0.022
σ^2_{FRow}	0.000	0.014
σ^2_{FCol}	0.013	0.025
σ^2_{HRange}	0.038	0.035
σ^2_{HRow}	0.092	0.088
ψ_{a_F}	0.000	0.023
ψ_{e_F}	2.458	1.617
λ_{a_H}	0.151	0.335
λ_{a_F}	-1.074	-0.032
λ_{e_H}	0.376	0.997
λ_{e_F}	-0.591	-0.874
ϕ_H	1.095	0.965
σ^2_F	3.593	3.571

σ^2_j, $j =$ Tank, Rack, FRow, FCol, HRange and HRow are the variance components associated with the peripheral random effects, ψ_{a_F}, ψ_{e_F}, λ_{a_H}, λ_{a_F}, λ_{e_H} and λ_{e_F} are the genetic parameters and ϕ_H and σ^2_F are the hydroponic dispersion and field residual variance parameter respectively

Conclusion

Conclusion

- This research question posed a range of complexities to address in order to provide a valid statistical analysis

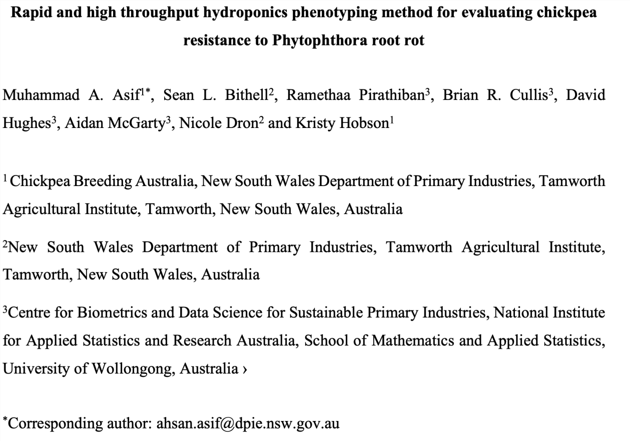
Conclusion

- This research question posed a range of complexities to address in order to provide a valid statistical analysis
- The resulting analysis gave a quantification of the level of agreement between PRR resistance for the two experiments

Conclusion

- This research question posed a range of complexities to address in order to provide a valid statistical analysis
- The resulting analysis gave a quantification of the level of agreement between PRR resistance for the two experiments
- Parametric bootstrapping provided assessment of parameter estimation bias and indicated the associated 95% bootstrap CI for the total genetic correlation does not contain 0

Paper



Rapid and high throughput hydroponics phenotyping method for evaluating chickpea resistance to Phytophthora root rot

Muhammad A. Asif[1*], Sean L. Bithell[2], Ramethaa Pirathiban[3], Brian R. Cullis[3], David Hughes[3], Aidan McGarty[3], Nicole Dron[2] and Kristy Hobson[1]

[1] Chickpea Breeding Australia, New South Wales Department of Primary Industries, Tamworth Agricultural Institute, Tamworth, New South Wales, Australia

[2] New South Wales Department of Primary Industries, Tamworth Agricultural Institute, Tamworth, New South Wales, Australia

[3] Centre for Biometrics and Data Science for Sustainable Primary Industries, National Institute for Applied Statistics and Research Australia, School of Mathematics and Applied Statistics, University of Wollongong, Australia ›

[*] Corresponding author: ahsan.asif@dpie.nsw.gov.au

References i

- Amritha Amalraj, Julian Taylor, Sean Bithell, Yongle Li, Kevin Moore, Kristy Hobson, and Tim Sutton.

  Mapping resistance to phytophthora root rot identifies independent loci from cultivated (cicer arietinum l.) and wild (cicer echinospermum p.h. davis) chickpea.

  Theoretical and Applied Genetics, 132:1017–1033, 4 2019.

- N R Bartlett.

  A survival model for a wood preservative trial, 1978.

References ii

N E Breslow and D G Clayton.

Approximate inference in generalised linear mixed models.

Journal of the American Statistical Association, 88:9–25, 1993.

N E Breslow and X Lin.

Bias correction in generalised linear mixed models with a single component of dispersion.

Biometrika, 82:81–91, 1995.

D Collins.

The performance of estimation methods for generalised linear mixed models, 2008.

References iii

Karin Meyer.

Reducing computational demands of restricted maximum likelihood estimation with genomic relationship matrices.

Genetics Selection Evolution, 55(1):7, 2023.