LegumeLegacy

LegumeLegacy SS6 Advanced Statistical Modelling

This project has received funding from the European Union's Horizon 2021 doctoral network programme under the Marie Skłodowska-Curie grant agreement No. 101072579.

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

This work has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI).

This work has received funding from UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number EP/X028003/1] to the University of Reading.

Racines d'avenir

The Canadian sites of this project were funded by the Living Laboratories Initiative (or Agricultural Climate Solutions – Living Labs) of Agriculture and Agri-Food Canada.

Modelling BEF relationships with repeated measurements over time using Diversity-Interactions models and the **`DImodelsMulti`** R package

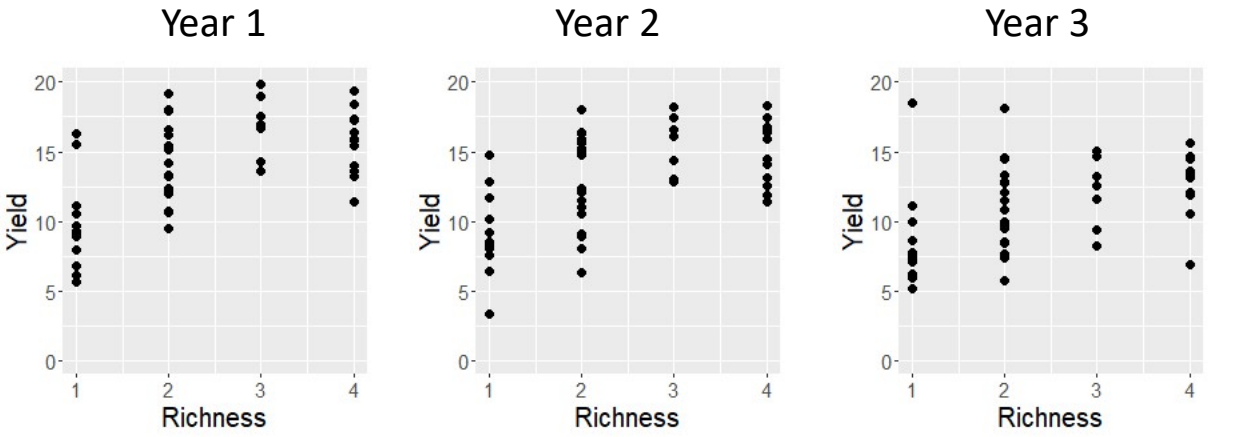
Caroline Brophy
Professor in Statistics
School of Computer Science and Statistics
Trinity College Dublin
Ireland

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

With thanks to Laura Byrne, Rafael Moral, James O'Malley and Rishabh Vishwakarma who developed some of the material in this presentation.

Biodiversity and Ecosystem Function data

Over time……



Analysing data over time?

- How to analyse?
  - Each time point can be analysed separately
  - All time points can be analysed jointly: repeated measures analysis
- Benefits of fitting a repeated measures model?
  - Allows testing for trends over time
  - Allows for comparisons between time points
- Downside of fitting a repeated measures model?
  - More complicated!
  - May not end up with the “best” model for any given time point

What is the difference mathematically?

Simple linear regression in matrix notation

The regular form of the SLR model is:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \text{where} \qquad \epsilon_i \sim \text{IID } N(0, \sigma^2)$$

where IID stands for independent and identically distributed.

This can be expressed as:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

where $\mathbf{Y}$ is the $n \times 1$ response vector, $\mathbf{X}$ is the $n \times 2$ design matrix, β is the 2×1 parameter vector, and ϵ is the $n \times 1$ error vector, and:

$$\text{E}(\epsilon) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \text{Var}(\epsilon) = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

Time: 1, 2, 3.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & & & \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & & 0 & \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & & & \\ & & & \ddots & & \\ & & & & \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ & 0 & & & \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ & & & & \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix}$$

Different error structures for three timepoints

Unstructured (UN)	Compound Symmetry (CS)	Autoregressive Model Order 1 (AR(1))
$\begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} \\ \sigma_{1,2} & \sigma_2^2 & \sigma_{2,3} \\ \sigma_{1,3} & \sigma_{2,3} & \sigma_3^2 \end{bmatrix}$	$\sigma^2 \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$	$\sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}$

What does a repeated measures model look like?

$$\begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

$$y = (\beta_0 + \beta_1 x) * \text{Time} + \epsilon$$

$$\epsilon \sim N(0, R)$$

For three times points:

$$R = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & & & \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & & 0 & \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & & & \\ & & & \ddots & & \\ & & & & \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ & 0 & & & \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ & & & & \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix}$$

Repeated measures model

- Account for covariances in the error structure
- Including time within the fixed effects. Options:
  - All fixed effects crossed with time (common)
  - Additive effect of time (main effect of time)
  - No fixed effect of time

Exercise

- The Diversity-Interactions 'ID' model is

$$y = \left(\sum_{i=1}^{s} \beta_i p_i\right) * \text{Time} + \epsilon \qquad \epsilon \sim N(0, \text{R})$$

- Write out the ID model in matrix notation for data from a BEF experiment where there are three species.

Dimensions of each matrix?

- The design consists of 15 plots each recorded over two time points:
  - Each species occurs in monoculture in two plots,
  - Each pair of species occurs in a 50:50 mixture in two plots,
  - The three species occurred in equal proportions in three plots.

Four-species example
With repeated measures

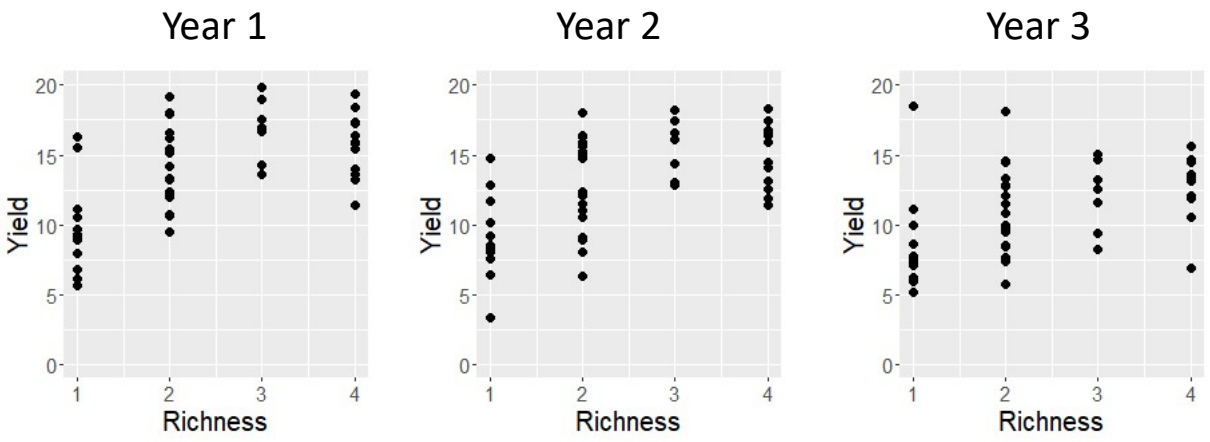
Experiment design

- Dataset name: Dataset_4species_rep_measures_simulated.csv
- Simulated assuming a biodiversity and ecosystem function (BEF) experiment
- Four grassland species (species 1 and 2 grasses, species 3 and 4 legumes), let's call them G1, G2, L1 and L2
- Species diversity manipulated across 50 plots
  - Monocultures
  - Mixtures: binary 50:50 and unbalanced, three species equi-proportional, four species equi-proportional
- Annual yield of each plot recorded over three growing seasons

In the dataset…

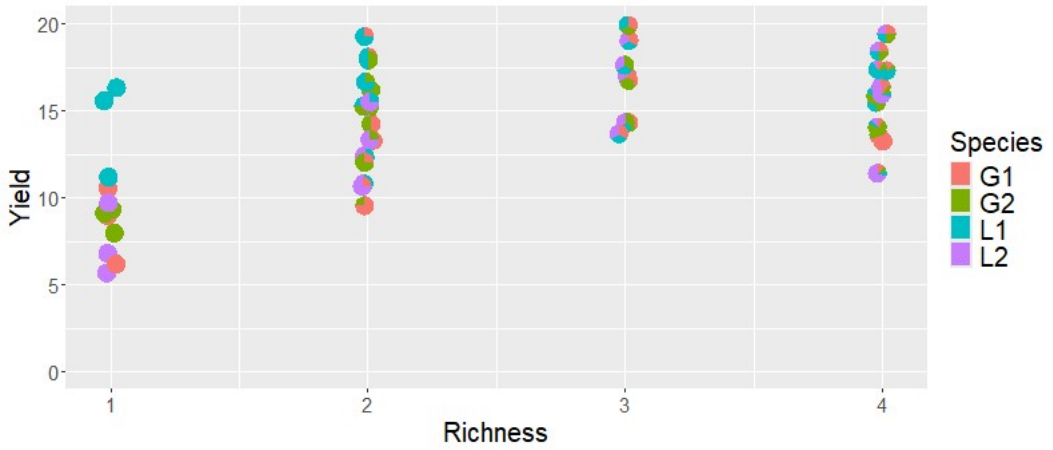
- Yearn: year (1, 2 or 3) that measurement was recorded
- Plot: unique ID for each plot
- Community: ID for each unique species diversity treatment
- Richness: number of species sown in a plot
- G1, G2, L1, L2: the proportion sown for each species
- Yield: the yield measured on each plot (at the given time point)

Raw data



Four-species example. Year 1 analysis.
Recap on analysing data from a single point in time

Year 1 data



Year 1 data – model selection process

- Using autoDI in the DImodels package:
  - Theta = 1
  - Functional group interaction structure
  - No evidence of lack of fit
- Additional model selection needed?
- Diagnostics?

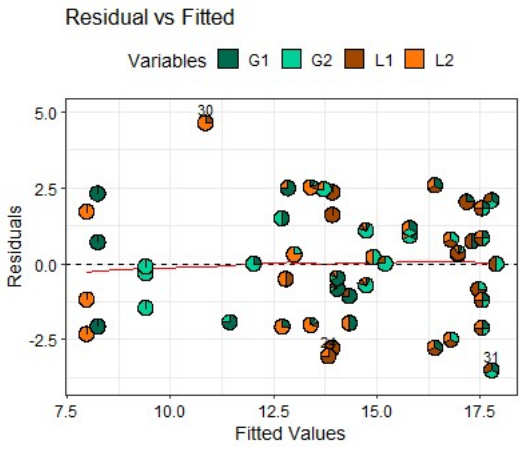
Estimated model

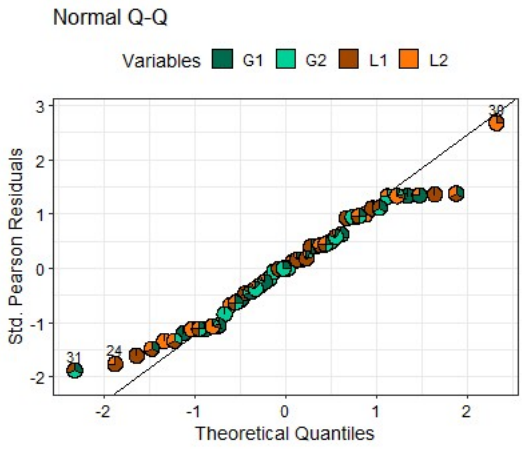
Term	Estimate	Std. Error	P-value
G1	8.24	0.9185	
G2	9.40	0.9185	
L1	13.92	0.9185	
L2	7.97	0.9185	
GL	24.93	2.7139	<0.004
GG	15.51	5.4621	0.007
LL	7.42	5.4621	0.182

How to make sense of this model…….?

Does the model fit well?

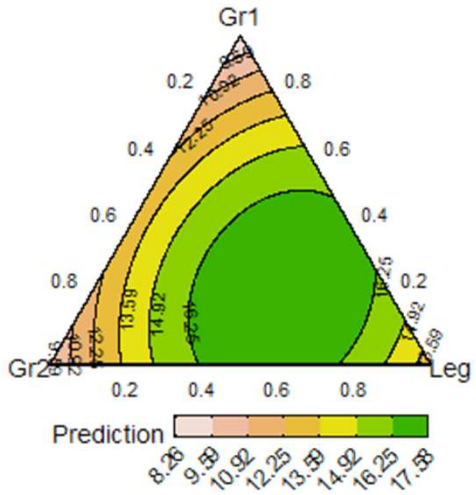
Term	Est	SE	P
G1	8.24	0.9185	
G2	9.40	0.9185	
L1	13.92	0.9185	
L2	7.97	0.9185	
GL	24.93	2.7139	<0.004
GG	15.51	5.4621	0.007
LL	7.42	5.4621	0.182





Interpreting the model

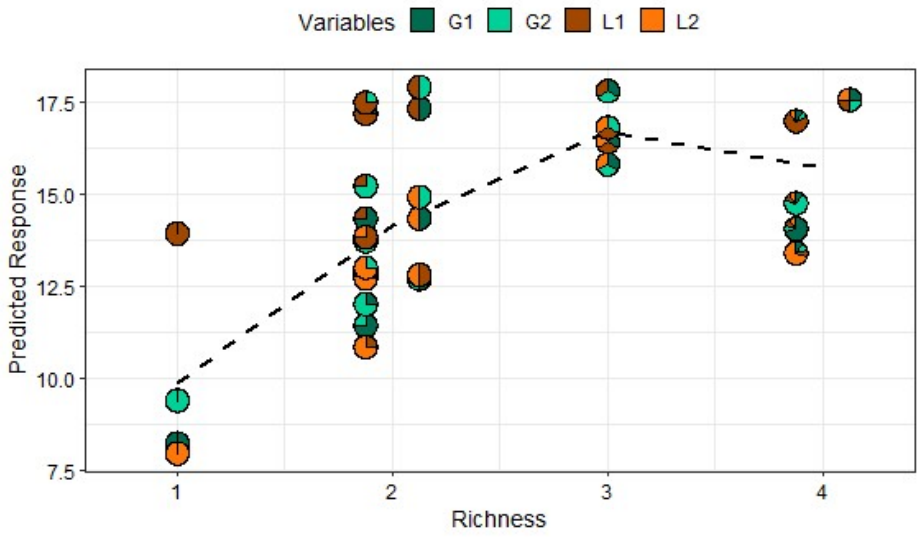
Term	Est	SE	P
G1	8.24	0.9185	
G2	9.40	0.9185	
L1	13.92	0.9185	
L2	7.97	0.9185	
GL	24.93	2.7139	<0.004
GG	15.51	5.4621	0.007
LL	7.42	5.4621	0.182



Two legume species in equal proportions at each point in the ternary

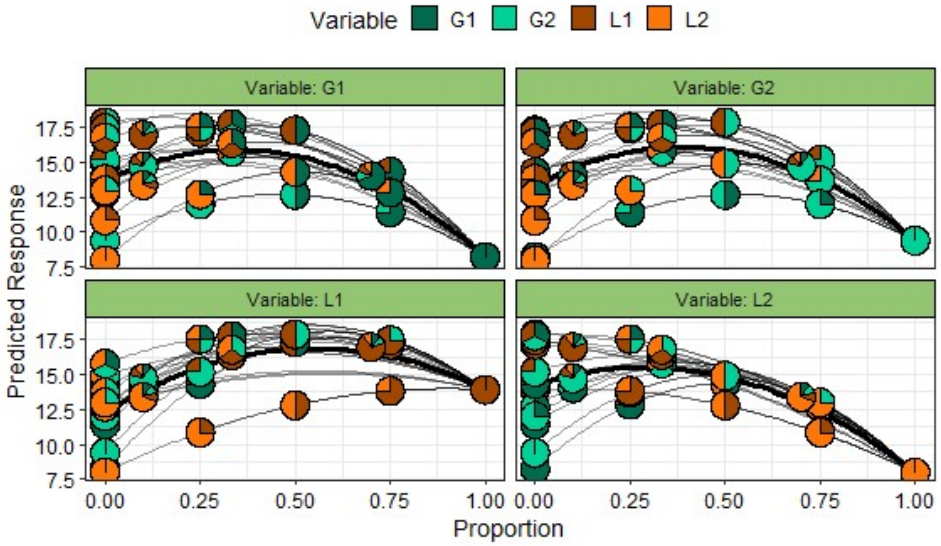
Interpreting the model

Term	Est	SE	P
G1	8.24	0.9185	
G2	9.40	0.9185	
L1	13.92	0.9185	
L2	7.97	0.9185	
GL	24.93	2.7139	<0.004
GG	15.51	5.4621	0.007
LL	7.42	5.4621	0.182



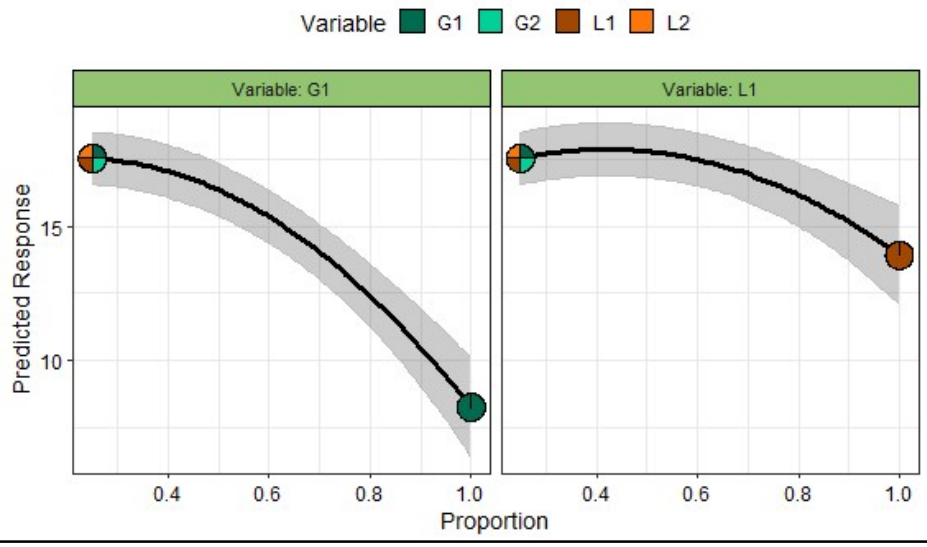
Interpreting the model

Term	Est	SE	P
G1	8.24	0.9185	
G2	9.40	0.9185	
L1	13.92	0.9185	
L2	7.97	0.9185	
GL	24.93	2.7139	<0.004
GG	15.51	5.4621	0.007
LL	7.42	5.4621	0.182



Interpreting the model

Term	Est	SE	P
G1	8.24	0.9185	
G2	9.40	0.9185	
L1	13.92	0.9185	
L2	7.97	0.9185	
GL	24.93	2.7139	<0.004
GG	15.51	5.4621	0.007
LL	7.42	5.4621	0.182



Four-species example. Analysis of all three years.
Repeated measures Diversity-Interactions model

Back to the dataset…

- Yearn: year (1, 2 or 3) that measurement was recorded
- Plot: unique ID for each plot
- Community: ID for each unique species diversity treatment
- Richness: number of species sown in a plot
- G1, G2, L1, L2: the proportion sown for each species
- Yield: the yield measured on each plot (at the given time point)

Fitting a DI model to repeated measures data

- Diversity-Interactions 'ID' model is

$$y = \left(\sum_{i=1}^{4} \beta_i p_i \quad + \text{Interactions} \right) * \text{Yearn} + \epsilon \qquad \epsilon \sim N(0, \text{R})$$

- Model selection
  - What is the best interaction structure?
  - What is the best error structure?

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & & & \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & & 0 & \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & & & \\ & & & \ddots & & \\ & & & & \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ & 0 & & & \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ & & & & \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix}$$

Model selection process

- Theta? Interaction structure? Error structure?

Important:

- When comparing fixed effects, use ML estimation
- When comparing error structures, use REML estimation
- When model selection is complete, fit the final model using REML estimation.

Model selection – theta?

- Using the DImodelsMulti R package:

```
DImulti(y = "yield", time = c("yearn", "UN"), unit_IDs = "plot",
prop = c("G1", "G2", "L1", "L2"), FG = c("G","G","L","L"),
data = dataRM, DImodel = "FULL",
     method = "ML", estimate_theta = FALSE)
```

```
estimate_theta = TRUE
```

ML estimation

Set theta equal to 1.

- With $\theta = 1$: AICc = 658.3
- With $\hat{\theta} = 0.93$: AICc = 657.9

Model selection – Interaction structure

	AICc
• ID model	695.3
• AV model	655.1
• FG model	638.4
• FULL model	658.3

Repeated measures model with UN (unstructured) covariance matrix, vary the interaction structures

Use FG interactions

ML estimation

```
DImulti(y = "yield", time = c("yearn", "UN"), unit_IDs = "plot",
        prop = c("G1", "G2", "L1", "L2"), FG = c("G","G","L","L"),
        data = dataRM, DImodel = "FG",
        method = "ML", estimate_theta = FALSE)
```

Model selection – variance covariance matrix

	AICc
• UN	587.4
• CS	604.0
• AR(1)	590.1

Repeated measures model with FG interaction structures, testing covariance structures

Use UN covariance structure

REML estimation

```
DImulti(y = "yield", time = c("yearn", "UN"), unit_IDs = "plot",
        prop = c("G1", "G2", "L1", "L2"), FG = c("G","G","L","L"),
        data = dataRM, DImodel = "FG",
        method = "REML", estimate_theta = FALSE)
```

Final model

All fixed effects crossed with year; error structure accounts for covariances between repeated measures.

REML estimation

	Year 1		Year 2		Year 3	
Term	Estimate	SE	Estimate	SE	Estimate	SE
G1	8.2	0.919	6.4	0.951	7.7	1.239
G2	9.4	0.919	10.5	0.951	8.8	1.239
L1	13.9	0.919	10.7	0.951	11.8	1.239
L2	8.0	0.919	8.1	0.951	6.5	1.239
GL	24.9	2.714	28.2	2.81	13.8	3.66
GG	15.5	5.463	11.2	5.656	11.8	7.366
LL	7.4	5.463	-3.6	5.656	3.8	7.366

Vcov matrix

$$\begin{bmatrix} 3.85 & 2.99 & 1.78 \\ 2.99 & 4.11 & 3.29 \\ 1.78 & 3.29 & 6.98 \end{bmatrix}$$

Comparison of communities

Does the grass 1 identity effect change from year 1 to 2?

Year 1: $\hat{\beta}_{11} = 8.24$ with standard error 0.919

Year 2: $\hat{\beta}_{12} = 6.36$ with standard error 0.951

Difference: 1.87, with standard error of difference 0.6570.

The grass identity effect declined from year 1 to year 2 ($p<0.01$)

Vcov matrix

$$\begin{bmatrix} 3.85 & 2.99 & 1.78 \\ 2.99 & 4.11 & 3.29 \\ 1.78 & 3.29 & 6.98 \end{bmatrix}$$

Feeds into estimated variance covariance matrix for the parameters estimates

Comparison of communities

Does the equi-proportional four-species mix change from year 1 to 2?

Year 1: $\hat{y} = 17.54$ with standard error

Year 2: $\hat{y} = 16.45$ with standard error

Difference: 1.09, with standard error of difference 0.3462.

The equi-proportional four-species mix declined from year 1 to year 2 ($p<0.01$)

Vcov matrix

$$\begin{bmatrix} 3.85 & 2.99 & 1.78 \\ 2.99 & 4.11 & 3.29 \\ 1.78 & 3.29 & 6.98 \end{bmatrix}$$

Feeds into estimated variance covariance matrix for the parameters estimates

Comparison of communities

Does the equi-proportional four-species mix change from year 2 to 3?

Year 1: $\hat{y} = 16.45$ with standard error

Year 2: $\hat{y} = 13.10$ with standard error

Difference: 3.35, with standard error of difference 0.5257.

The equi-proportional four-species mix declined further from year 2 to year 3 (p<0.001)

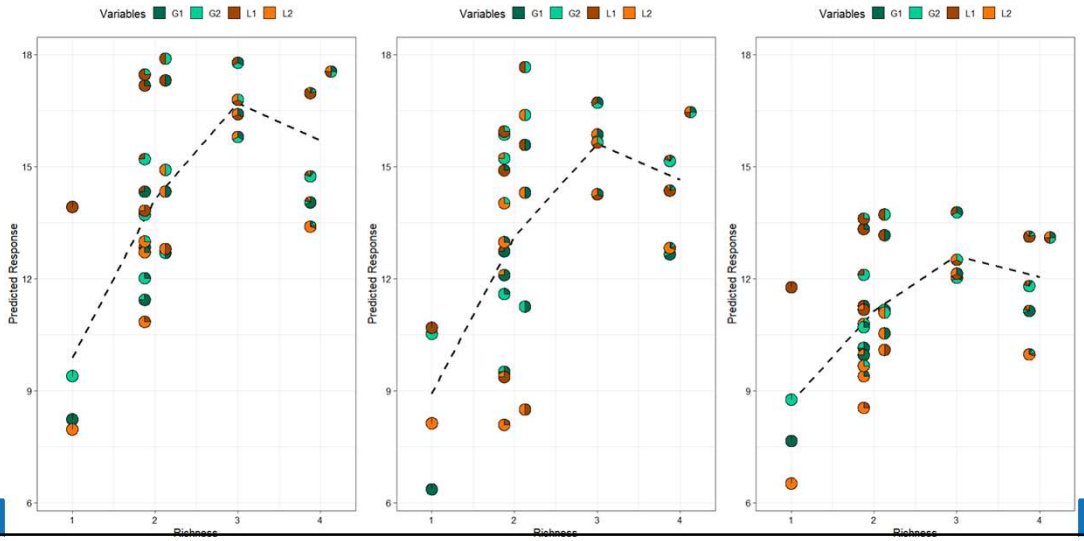
Vcov matrix

$$\begin{bmatrix} 3.85 & 2.99 & 1.78 \\ 2.99 & 4.11 & 3.29 \\ 1.78 & 3.29 & 6.98 \end{bmatrix}$$

Feeds into estimated variance covariance matrix for the parameters estimates

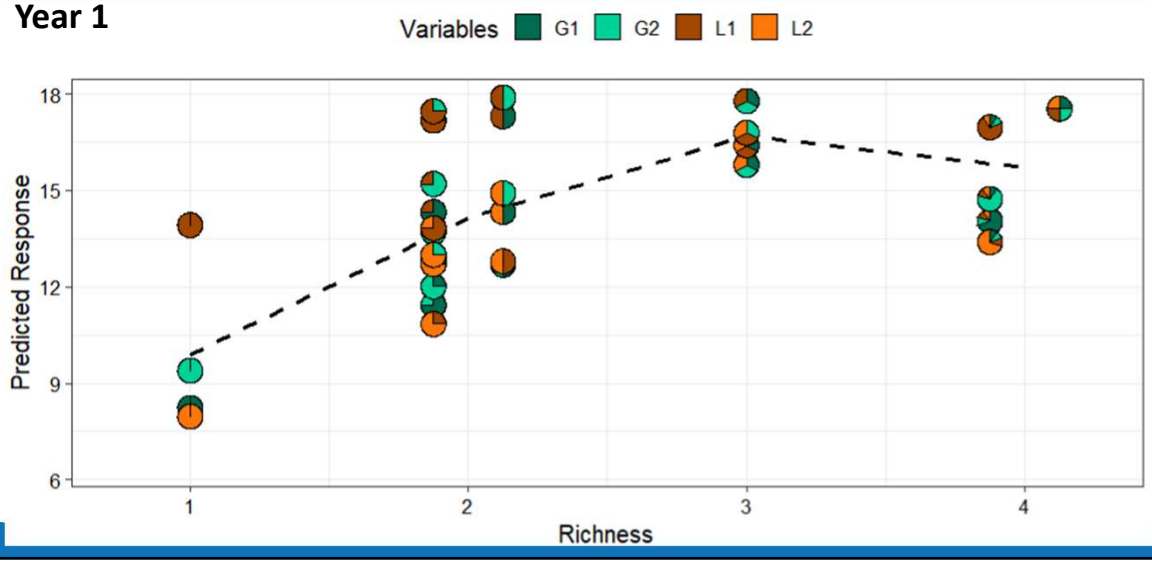
Model predictions as richness changes

Year 1 **Year 2** **Year 3**



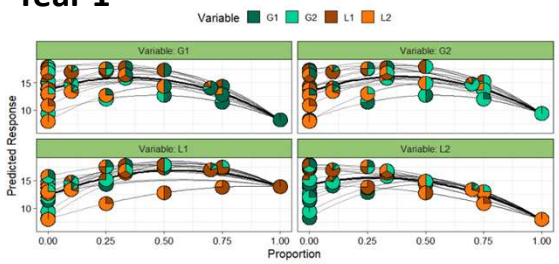
Model predictions as richness changes

Year 1

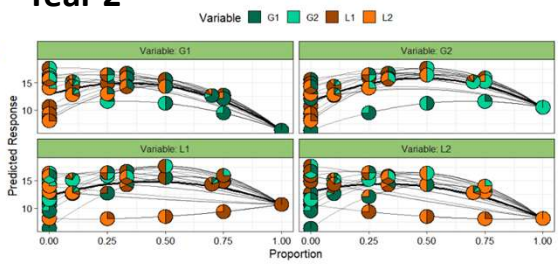


Predictions across changing species proportions

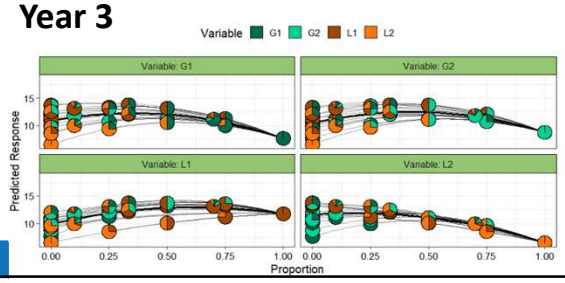
Year 1



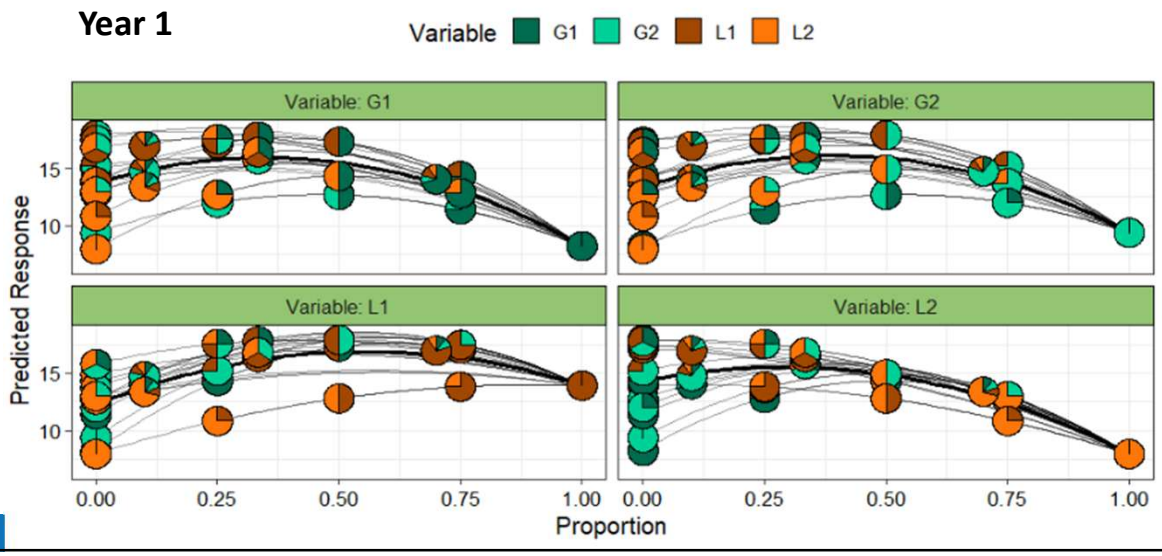
Year 2



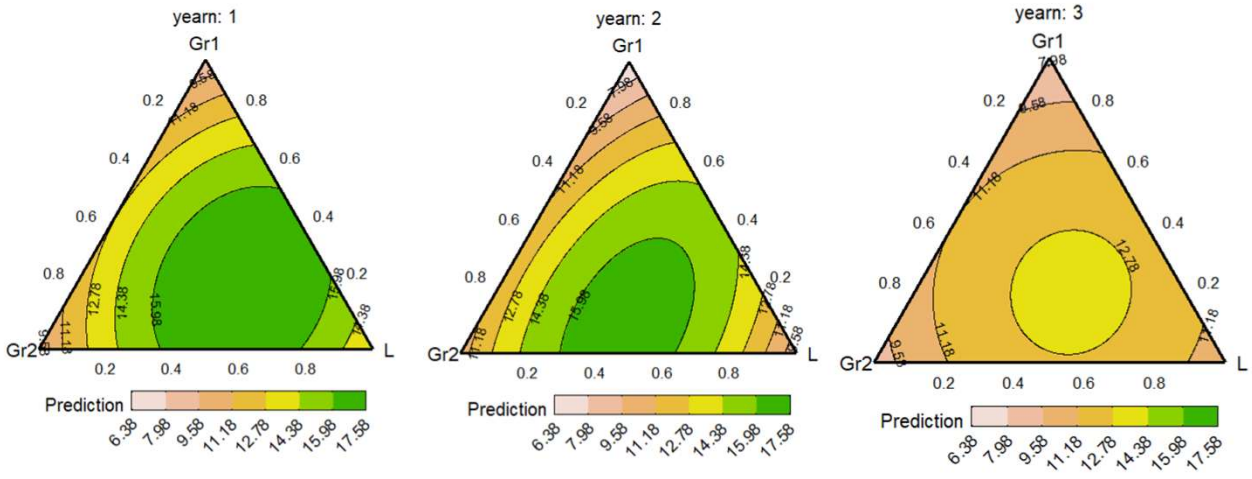
Year 3



Predictions across changing species proportions



Predicting at G1, G2, L combinations



In summary…

- Do I need to use repeated measures analysis when I have data over time?
  - Maybe…
  - Maybe not….

What are the research questions?

- If modelling data over time all together, crucial to build in covariances otherwise the analysis would not be valid and you would have pseudo-replication.