Linear Models in R

Helene Wagner, University of Toronto

Goal: Test for adaptation in Western white pine

- Estimate trait from common garden experiment
- Assess trait heritability
- Compare trait variation to neutral markers (SNPs)
- Explain trait variation by 'climate' and 'geography'

Trait: d13C
(related to water use efficieny)

Image source: wikipedia.com

Methodological Challenges

1. Design of Western white pine study
2. Introducing SNP data
3. Specifying linear models
   - What type of linear model to fit?
   - Model formulas in R
4. Video 2: Under the hood
   - Degrees of freedom
   - Beyond p-values: effect size

Residual analysis (Week 4) Spatial linear models (Week 7) Model selection (Week 12)

Study Design

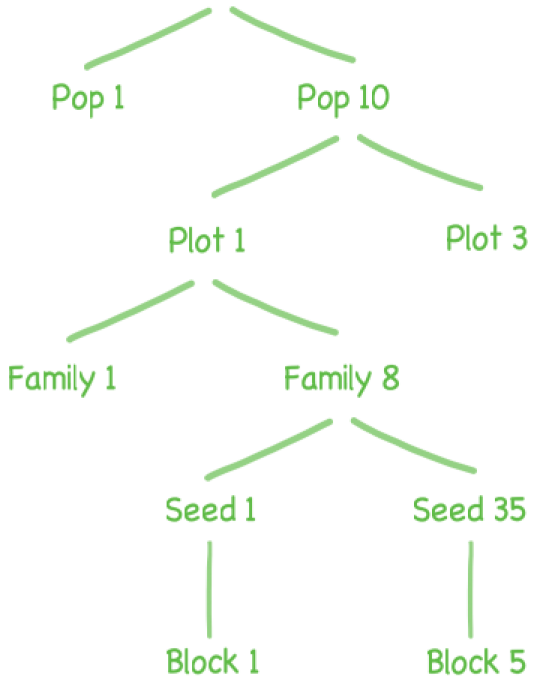
10 Populations

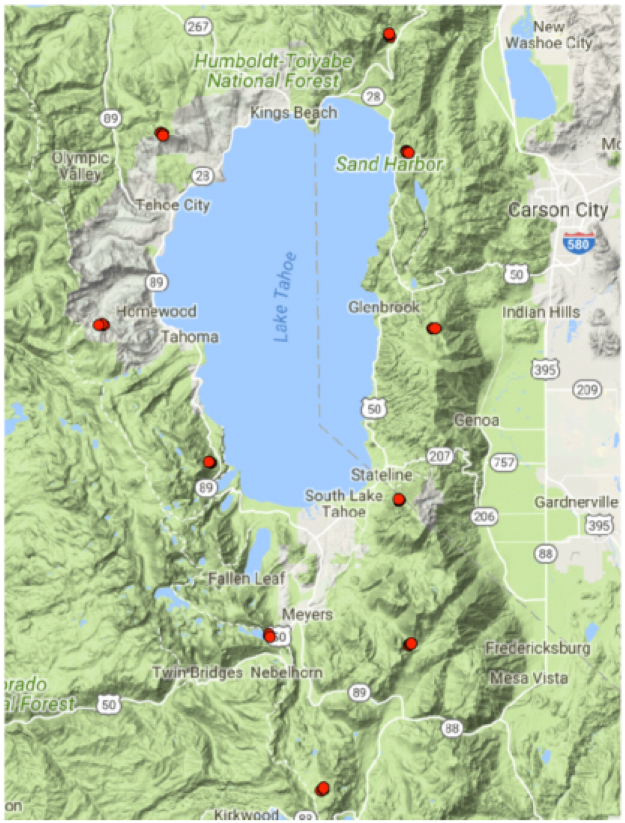
3 Plots per Population
- Spatial coordinates
- 7 bioclimatic variables

2 - 8 Trees per Plot
- 164 SNP markers

5 - 35 Seeds per Family

5 Blocks in Common Garden
- d13C of each seedling





Seedling trait:
- 3 - 6 seedlings per family
- common garden
- trait: d13C
- no molecular markers

Estimate heritable trait:
- Trait = G + E + G:E
- no trait measurement for mother
- estimate from offspring trait
- account for hierarchical sampling
- account for blocks

Genetic differentiation:
- Trait: calculate Qst
- SNPs: calculate Fst
- Compare Qst to Fst

Trait - Environment Association: Variation in heritable trait explained by Climate or Geography?

Introducing SNP Data

family <int>	population <fctr>	snp102.Plmn <fctr>	snp106.Plmn <fctr>
59	blk cyn	CC	AC
60	blk cyn	AC	AA
61	blk cyn	CC	AA
63	blk cyn	CC	AA
64	blk cyn	AC	AA
65	blk cyn	AC	AA
67	blk cyn	AC	AA
69	blk cyn	CC	AA
72	blk cyn	AC	AA
73	blk cyn	AC	*NA*

1–10 of 157 rows... Previous 1 2 3 4 5 6 ... 16

SNP markers:

- co-dominant
- genotype or haplotype?
- wildtype vs. mutant?
- monomorphic?
- many markers!
- neutral or not?

Here:

- 164 SNP markers
- monomorphic: 4
- 0 - 20% missing values
- unrelated to trait of interest

What Type of Linear Model to Fit?

1. Variable types?

Predictors X:	Response Y: Quantitative	Response Y: Binary
Quantitative	Regression (dummy vars)	Logistic Regression
Binary (single)	t - Test	Odds Ratio
Categorical	ANOVA (covariates)	Log-linear Model

2. LM, LMM, or GLMM?

Method:	Conditions: Normality	Conditions: Balanced
LM	✓	✓
LMM	✓	✗
GLMM	✗	✗

Example:
- Y: seedling d13C
- Hierarchical sampling: Families = trees within plots within populations
- Blocking: common garden with 5 blocks
- Covariate (plot level): AWS (soil available water supply)

3. Nested or crossed?

Nested factors: hierarchical sampling

Crossed factors: randomized experiment

	Block 1	Block 2	Block 3
Family 1	1	1	1
Family 2	1	1	~~1~~ 2
Family 3	1	1	~~1~~ 0

balanced unbalanced

4. Fixed or random factors?

Warning: Different philosophies!

	Replicate study	# Levels
Random	Different levels	Many (> 5?)
Fixed	Same levels	Often few

- Fixed effect: fits one parameter per level
- Random effect: fits 2 parameters

5. Interactions?

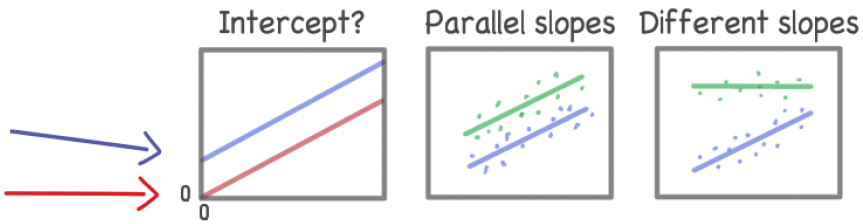
Two crossed factors: Pop, Block
- Effect of block (mean) depends on population (level of second factor)

Factor and covariate: Pop, AWS
- AWS has same effect for each pop (same slope, different intercepts)
- AWS has different effect (slope, intercept) for different pops.

Specify Model Formula in R

Factors only:	Fixed effects only:	With random effects:
- Global mean:	`d13C ~ 1`	
- Population mean:	`d13C ~ Pop`	`d13C ~ (1 | Pop)`
- Additive effects:	`d13C ~ Pop + Block`	`d13C ~ (1 | Pop) + (1 | Block)`
- Interaction:	`d13C ~ Pop * Block`	
- Hierarchical model:	`d13C ~ Pop/Family`	`d13C ~ (1 | Pop/Family)`

With covariate (AWS, standardized):



- Simple regression:	`d13C ~ AWS`	
- No intercept:	`d13C ~ -1 + AWS`	
- Parallel slopes:	`d13C ~ AWS + Block`	`d13C ~ AWS + (1 | Block)`
- Different slopes:	`d13C ~ AWS * Block`	`d13C ~ AWS + (AWS | Block)`

Interaction	Nested	Intercept	Sum	Quadratic
`A * B = A + B + A:B`	`A / B = A + A:B`	`A = 1 + A`	`I(A + B)`	`I(A^2)`

Degrees of Freedom

ANOVA F-test (LM)

F follows F-distribution with df(model), df(residual)

$$F = \frac{\textit{Variability between groups}}{\textit{Variability within groups}} = 1$$

$$F = \frac{\textit{Variability between groups}}{\textit{Variability within groups}} \gg 1$$

Assumptions

- Same variability in all groups
- Same group size (balanced)
- Residuals normally distributed
- Observations are independent

Degrees of freedom

df(total) = (n – 1)

- n = # independent observations
- one df used for global mean

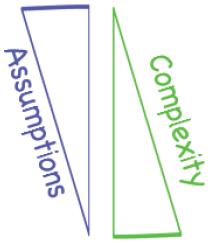
df(model) = (k – 1)

- k = number of parameters fitted

df(residual) = n – k

Alternative tests?

- Wald chi-square test
- Likelihood ratio test
- Conditional F-tests (balanced)
- Cond. F-tests with df correction
- MCMC or bootstrap tests



Non-independence?

Study design:

- Hierarchical sampling
- Blocking
- Paired samples
- Repeated measures

Autocorrelation:

- Spatial
- Temporal
- Phylogenetic

Co-ancestry:

- Kinship
- Population history
- Phylogeography

Beyond p-values: Effect Size

$$Eta^2 = \frac{Variability\ between\ groups}{Total\ variability} = \frac{Explained\ variance}{Total\ variance} = R^2$$

Variance Components

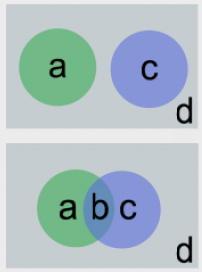
How much does each factor explain?

$$var\,(y) = \frac{\Sigma\,(y_i - \bar{y})^2}{(n-1)} = \frac{SS}{df_{total}}$$

	SS	Eta-squared
Factor A	0.35	0.35/0.9 = 0.39
Residual	0.55	0.55/0.9 = 0.61
Total	0.90	1.00

Covariates

What if predictors are correlated?



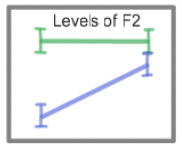
Crossed factors, balanced:
- Predictors uncorrelated
- No overlap of SS

Correlated predictors:
- Overlap of SS
- Results depend on model

Interactions

Which effects to interpret?

F1:F2 significant



No interaction

If 'F1 : F2' significant, don't interpret F1, F2!

Mixed Models

Avoid bias in variance estimates!

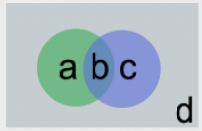
	Model Comparison	Model Fitting
LM	LS	LS
LMM	ML	REML

Best model: lowest AIC

Best predictor: highest SS

Variation Partitioning

Shared vs. unique explained variance



a: Climate alone (7 vars)
b: Shared variance
c: Geography alone (2 vars)

Partial regression (accounting for Z):
- Take residuals of Y ~ Z
- Take residuals of X ~ Z
- Fit model with residuals on both sides

Type I, II and III SS

The order of effects matters!

Type I: for nested factors?
Type II: balanced, additive?
Type III: crossed, interaction?

	Type I	Type II	Type III
F1	a+d+f+g	a+f	a
F2	b+e	b+e	b
F1:F2	c	c	c