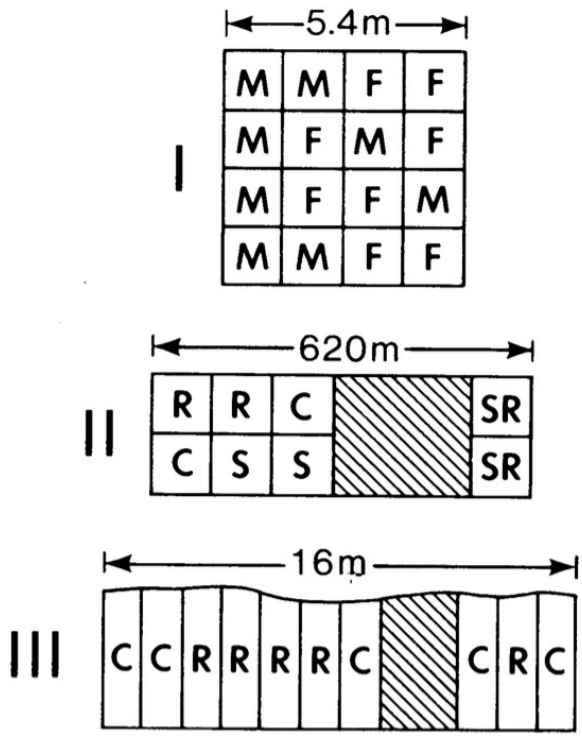
Coping with Additional Sources of Variation: ANCOVA and Random Effects

More Noise in Experiments & Observations

- Your 'fixed' coefficients are not always so fixed
- Continuous variation between samples can influence results
- Or samples may be non-independent - or pseudoreplicated!
- How do we deal with these problems in analyses?

ANCOVA and Quantified Variation

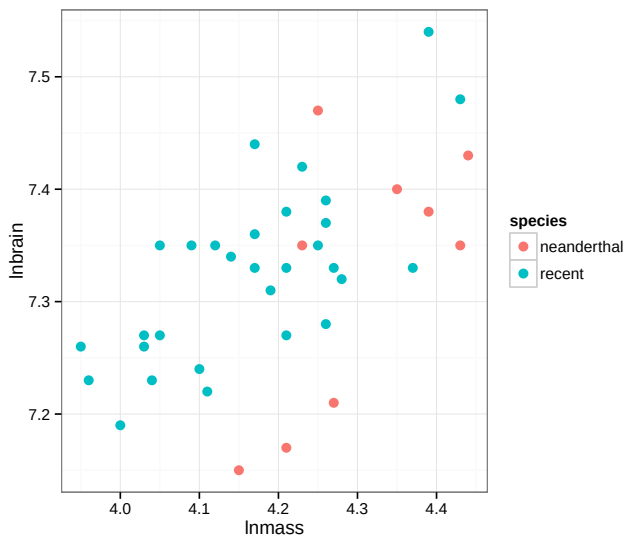
Gradients Bring Quantifiable Additional Variation



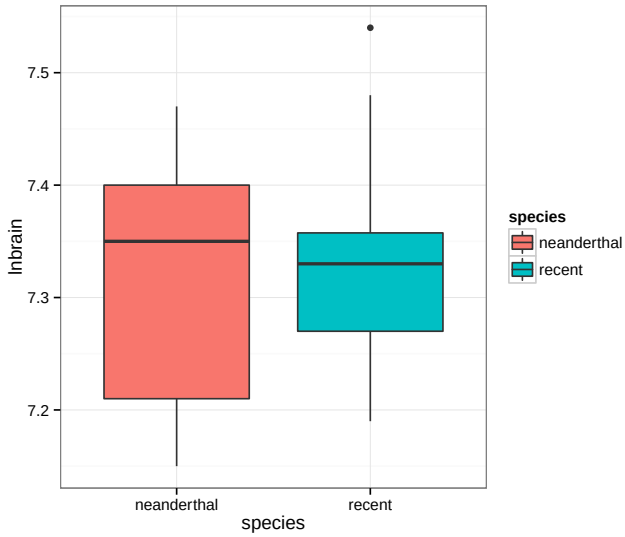
Neanderthals and the General Linear Model

How big was their brain?

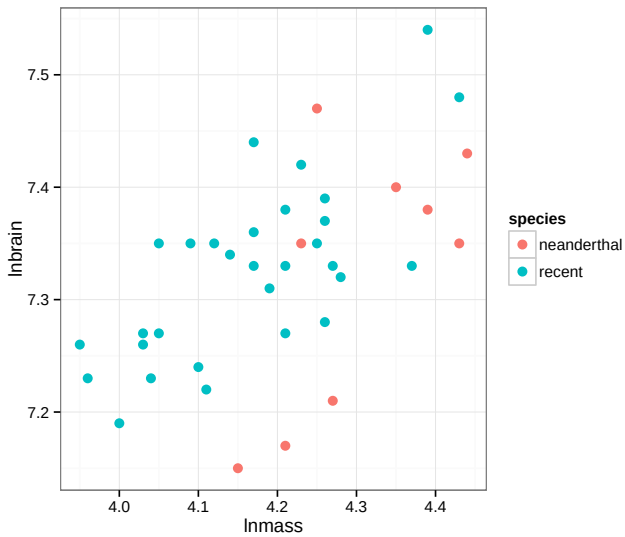
Problem: How Do you Evalute a Categorical Predictor in the Presence of a Continuous Predictor?



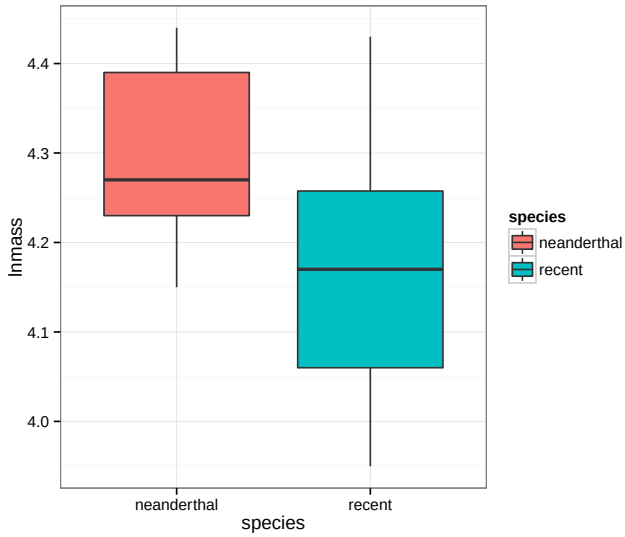
The Means Look the Same...



But there appears to be a Relationship Between Body and Brain Mass



And Mean Body Mass is Different

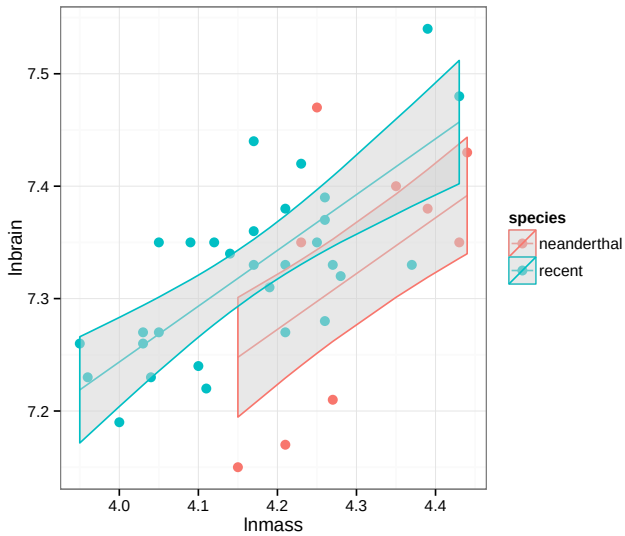


The General Linear Model

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

- This equation is huge. X can be anything - categorical, continuous, etc.
- We can control for the effect of a covariate - i.e., ANCOVA
- Type of SS matters, as 'covariate' is de facto 'unbalanced'
- With ANCOVA, we typically test for an interaction first, and if it is missing, drop it for better parameter estimation

Analysis of Covariance (control for a covariate)

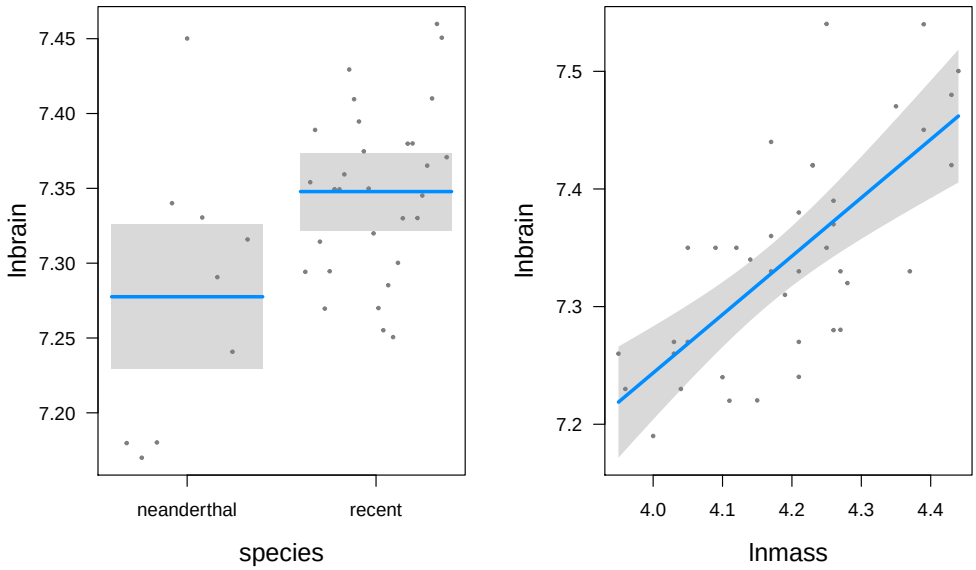


ANCOVA: Evaluate a categorical effect(s), controlling for a *covariate* (parallel lines)
Groups modify the *intercept*.

Exercise: Fit like a cave man

- Fit a model that will describe brain size from this data
- Does species matter? Compare type I and type II SS results
- Is there an interaction between size and species?
- Use Component-Residual plots or visreg to evaluate results

Species Effect Visually



Species Effect: Coefficients

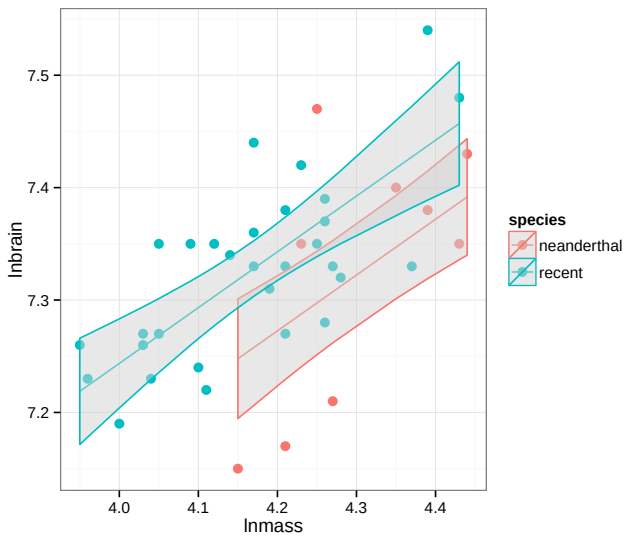
```
summary(neand_lm)$coefficients

#                Estimate Std. Error   t value     Pr(>|t|)
# (Intercept)   5.1880712 0.39525702 13.125817 2.736175e-15
# speciesrecent 0.0702784 0.02821518  2.490802 1.749474e-02
# lnmass        0.4963161 0.09172755  5.410764 4.262040e-06

summary(neand_lm)$r.squared

# [1] 0.4486172
```

Species Effect: Posthoc Test

```
contrast(neand_lm,
         list(species="neanderthal", lnmass=mean(neand$lnmass)),
         list(species="recent", lnmass=mean(neand$lnmass)),
         type="average")

# lm model parameter contrast
#
#     Contrast        S.E.       Lower        Upper     t df
# 1 -0.0702784 0.02821518 -0.1275014 -0.01305537 -2.49 36
#   Pr(>|t|)
# 1   0.0175
```

For more detail, try the `lsmeans` package

How to Plot a Fit Model



How to Properly Plot a Fit Model

```
neand <- cbind(neand, predict(neand_lm, interval="confidence"))

neand_plot +
  geom_line(data=neand, aes(y=fit)) +
  geom_ribbon(data=neand, aes(ymin=lwr,
                              ymax=upr),
              fill="lightgrey",
              alpha=0.5)
```


Random Effects

Moving into Modeling Variance

So far we have be fitting

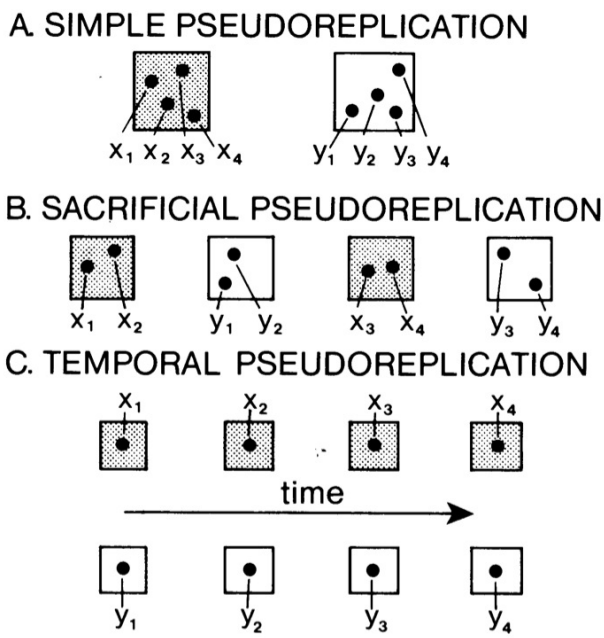
$$y_i = \beta_i X + \epsilon_i$$

where X is a number of predictors and $epsilon$ is random variation due to other processes. We assume data points are independent. But what if they're not? What if clusters of data points vary due to some random variation unique to just those points. We need a new model. One where

$$y_i = \alpha_{j[i]} + \beta_i X + \epsilon_{ij}$$

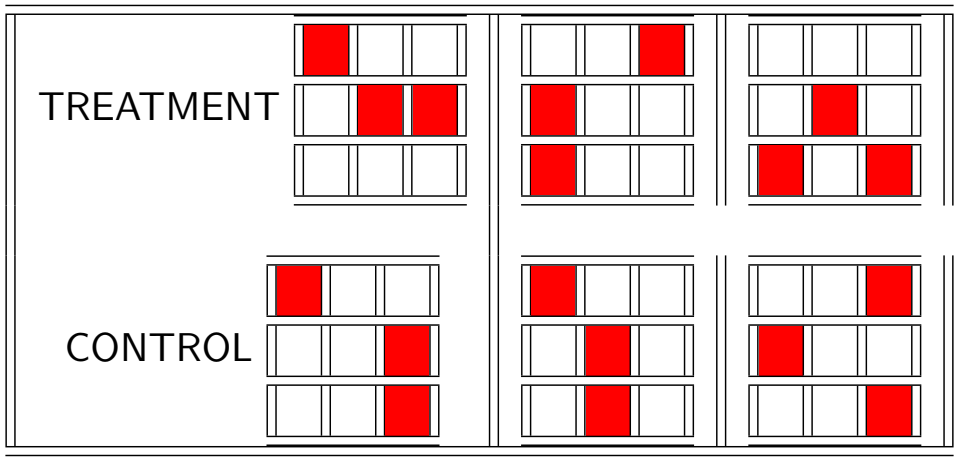
where i = individual data points, j = cluster, or group

This Framework Addresses Psuedoreplication Naturally



(Hurlbert 1984)

For Example, the Nested Design



Examples of Nesting

- Plots in a field with 1 treatment each
- Sampling a subject over time (where time doesn't influence the response)
- Gender of individuals (individual nested in gender)
- Experimental units manipulated by the same machine

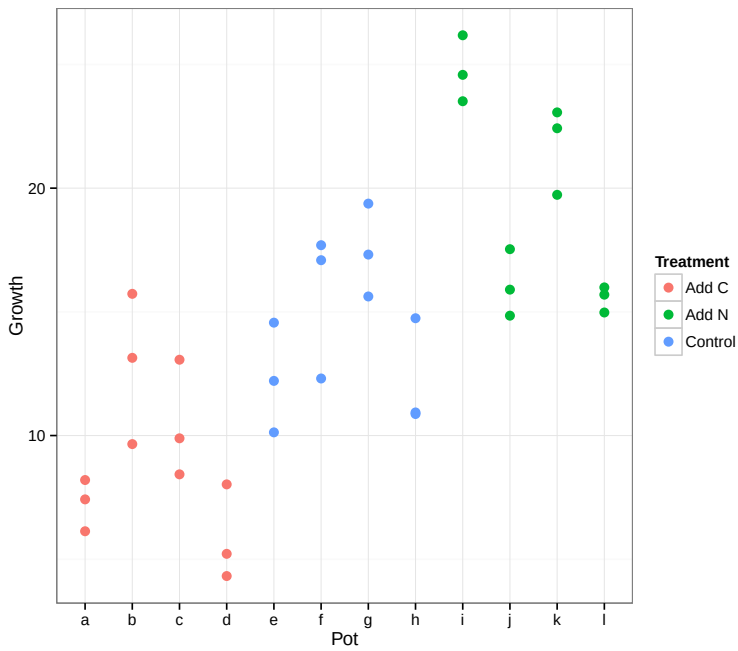
A Greenhouse Experiment testing C:N Ratios

Sam was testing how changing the C:N Ratio of soil affected plant leaf growth. He had 3 treatments. A control, a C addition, and a N addition. To ensure that any one measurement of one leaf wasn't a fluke, Sam measured 3 leaves per plant. The design is as follows:

3 Treatments (Control, C, N)
4 Pots of Plants per Treatment
3 Leaves Measured Per Pot

1) How many replicates are there per treatment?
2) Are measurements independent?
3) What do we use for our denominator Mean Square for F Test?
4) What is the denominator degrees of freedom?

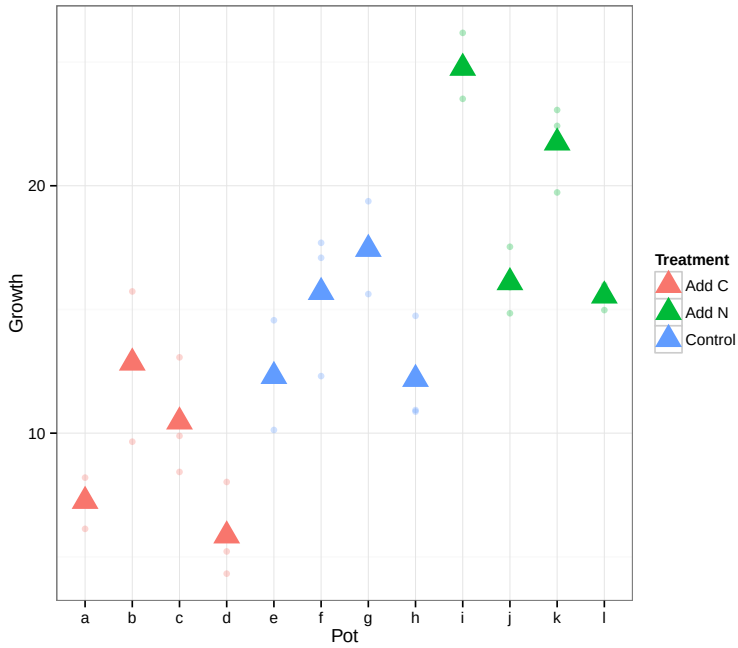
A Greenhouse Experiment testing C:N Ratios



Data Points are Not Independent!

Option 1: Averaging

If your design is balanced, and you don't care about the within pot variance, just average within each pot.



Option 1: Averaging

```
# Anova Table (Type II tests)
#
# Response: Growth
#           Sum Sq Df F value   Pr(>F)
# Treatment 217.68  2  8.9158 0.007331
# Residuals 109.87  9
```

You can use residuals to evaluate within plot variation.

Option 2: Classical ANOVA Error Decomposition with Expected Mean Squares

$$SS_{Total} = SS_{Treatment} + SS_{PotError} + SS_{WithinPotError}$$

```
plantAOV <- aov(Growth ~ Treatment + Error(Pot), data=plants)
```

```
summary(plantAOV)

#
# Error: Pot
#           Df Sum Sq Mean Sq F value  Pr(>F)
# Treatment  2  653.0   326.5   8.916 0.00733
# Residuals  9  329.6    36.6
#
# Error: Within
#           Df Sum Sq Mean Sq F value Pr(>F)
# Residuals 24  97.69    4.07
```


Option 3: Hierarchical/Mixed Models

Option 3: Multilevel/Clustered/Hierarchical/Mixed Model

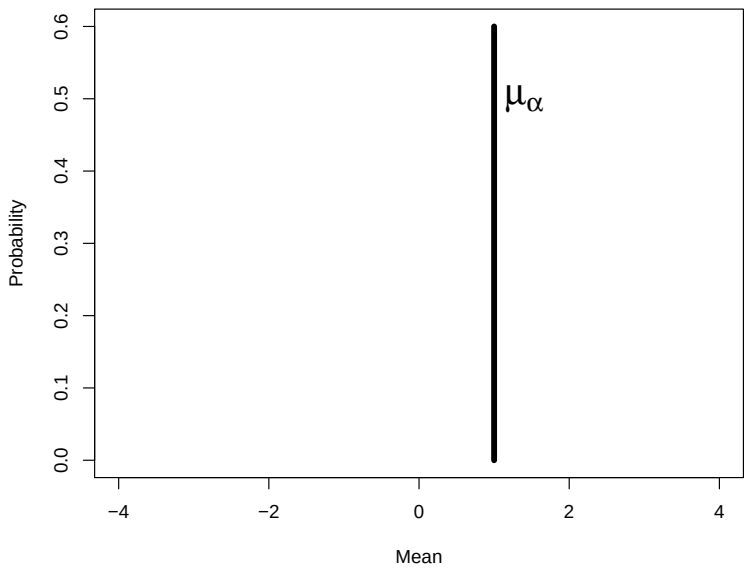
i = treatment, j = cluster, k = subsample

$$y_{ijk} = \alpha_{j[i]} + \beta_i X + \epsilon_{ijk}$$

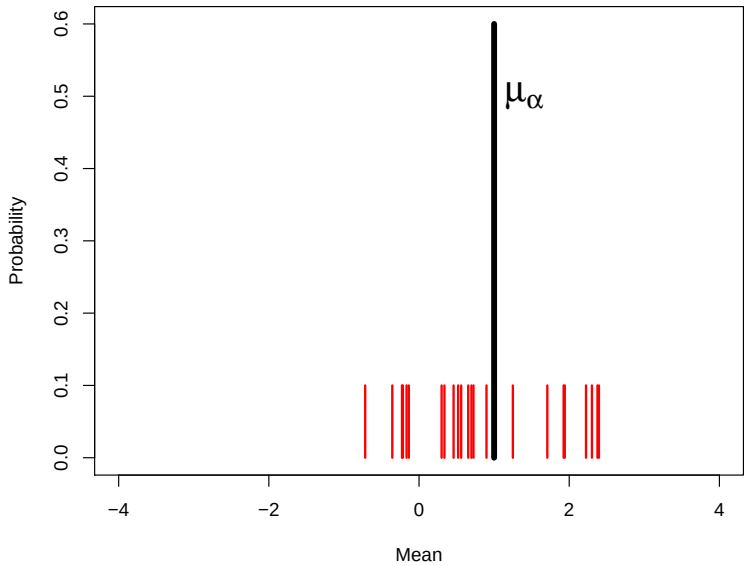
$$\alpha_{j[i]} \sim N(\mu_\alpha, \sigma^2_\alpha)$$

$$\epsilon_{ijk} \sim N(0, \sigma^2)$$

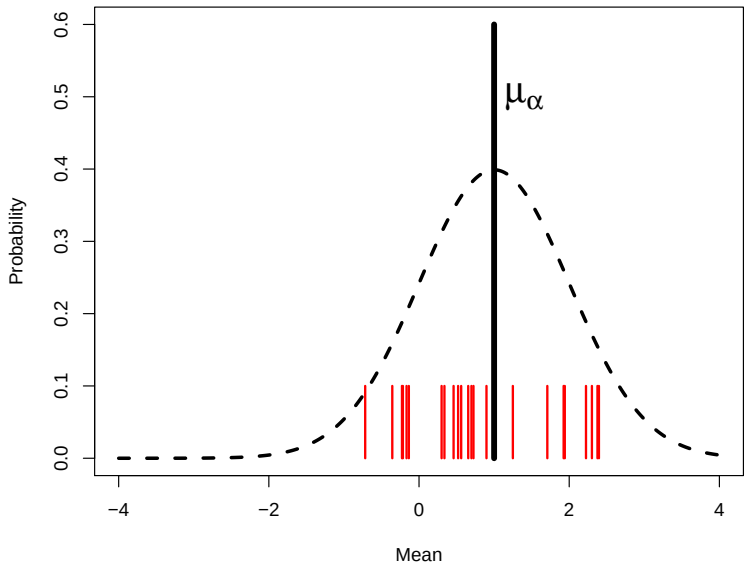
What is a random effect?



Each Plot has a Different Value of a Parameter



Values are Normally Distributed



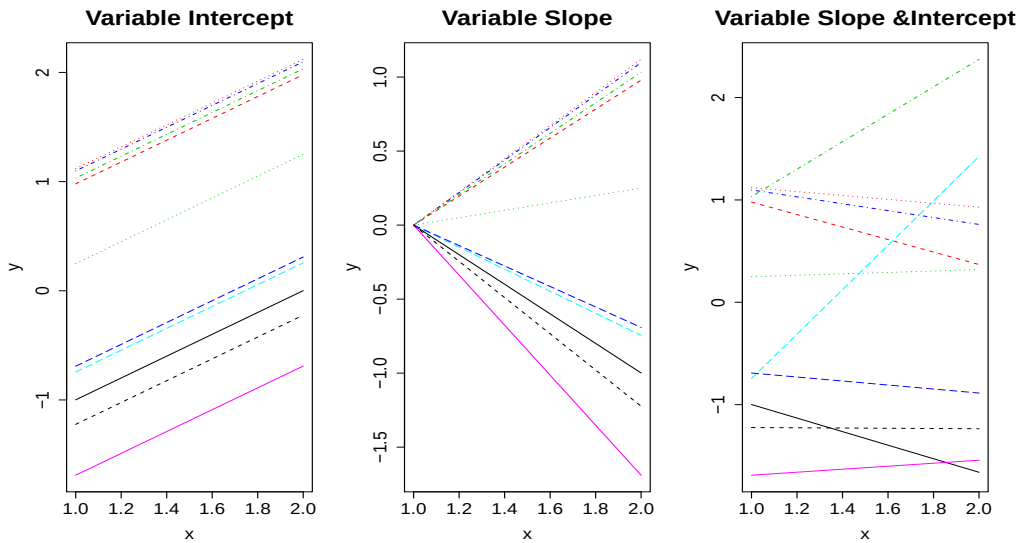
Types of Multilevel Models

i = treatment, j = cluster, k = subsample
Varying Intercept: $y_{ijk} = \alpha_{j[i]} + \beta_i X + \epsilon_{ijk}$
$\alpha_{j[i]} \sim N(\mu_\alpha, \sigma^2_\alpha)$

Varying Slope: $y_{ijk} = \alpha + \beta_{i[j]} X + \epsilon_{ijk}$
$\beta_{j[i]} \sim N(\mu_\beta, \sigma^2_\beta)$

Varying Slope & Intercept: $y_{ijk} = \alpha_{j[i]} + \beta_{j[i]} X + \epsilon_{ijk}$
$$\begin{pmatrix} \alpha_{j[i]} \\ \beta_{j[i]} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \sigma^2_\alpha & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma^2_\beta \end{pmatrix} \right)$$

Types of Multilevel Models



Unlike the General Linear Model, slopes and intercepts are constrained around a normal distribution

Fixed versus Random Effects

Fixed Effect: Effects that are constant across populations.
Random Effect: Effects that vary are random outcomes of underlying processes.

Gelman and Hill (2007) see the distinction as artificial. Fixed effects are special cases of random effects where the variance is infinite. The model is what you should focus on.

You will also hear that 'random effects' are effects with many levels, but that you have not sampled all of them, wheras a for fixed effects, you have sampled across the entire range of variation. This is subtly different, and artificial.

Some Points about Multilevel Models

- Flexible. Can accomodate varying slope, intercept, intercept-slope models
- Solved using Restricted Maximum Likelihood (REML). ML estimation produces downward biased estimates of random effect variances.
- As group level effects are drawn from the same distribution, Best Linear Unbiased Predictors (BLUPs) are shrunk towards grand mean - basically, we use information from all groups to inform within group means - useful for unbalanced designs.
- We will use one formulation to evaluate DF for p values, etc., but this is an ongoing research topic.

This is a BIG Topic, We are Diving Skin Deep

Useful Texts (which may show up in the future)

- Gelman A, Hill J, (2006) Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press
- Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith G (2009) Mixed Effects Models and Extensions in Ecology with R. Springer, New York.
- Pinheiro J, Bates D (2000) Mixed Effects Models in S and S-Plus. Springer-Verlag, New York, USA.

This is a BIG Topic, We are Diving Skin Deep

Websites that Discuss Mixed Models Regularly

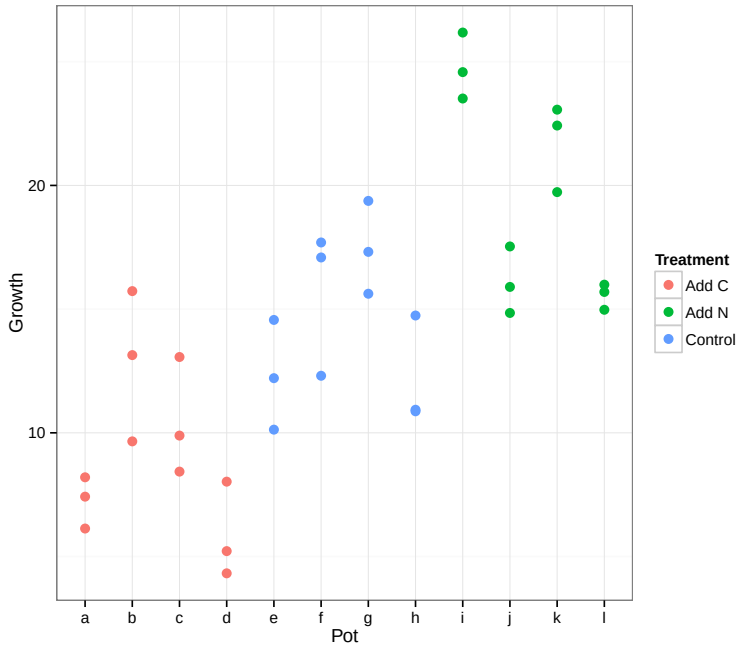
- `http://www.quantumforest.com/`
- `http://andrewgelman.com`
- `https://stat.ethz.ch/mailman/listinfo/r-sig-mixed-models`
- `http://glmm.wikidot.com/`

Many R Packages for Multilevel Models

- nlme - from Pinhero and Bates 2009
- lmer - bleeding edge by Doug Bates
- MCMCglmm - uses Bayesian techniques & MCMC (similar syntax to nlme)
- glmmADMB - interface for AD Model Builder

Fitting & Getting Results from Mixed Models

Back to A Greenhouse Experiment testing C:N Ratios



Leaf Growth = Treatment Effect + Pot Variation + Error

Fitting a Varying Intercept Model for the Greenhouse Experiment

```
library(nlme)
plantLME <- lme(Growth ~ Treatment, random = ~ 1|Pot, data=plants)
```

Summary: Random Effects

```
summary(plantLME)

# Linear mixed-effects model fit by REML
#  Data: plants
#        AIC      BIC    logLik
#   177.2003 184.6829 -83.60017
#
# Random effects:
#  Formula: ~1 | Pot
#         (Intercept) Residual
# StdDev:    3.294019 2.017524
#
# Fixed effects: Growth ~ Treatment
....
```


Summary: T-Tests for Fixed Effects

```
....

#                      Value Std.Error DF  t-value p-value
# (Intercept)       9.102532  1.746952 24 5.210523  0.0000
# TreatmentAdd N   10.432122  2.470563  9 4.222569  0.0022
# TreatmentControl  5.301297  2.470563  9 2.145785  0.0604
....
```

Fixed Effects v. Net Coefficients

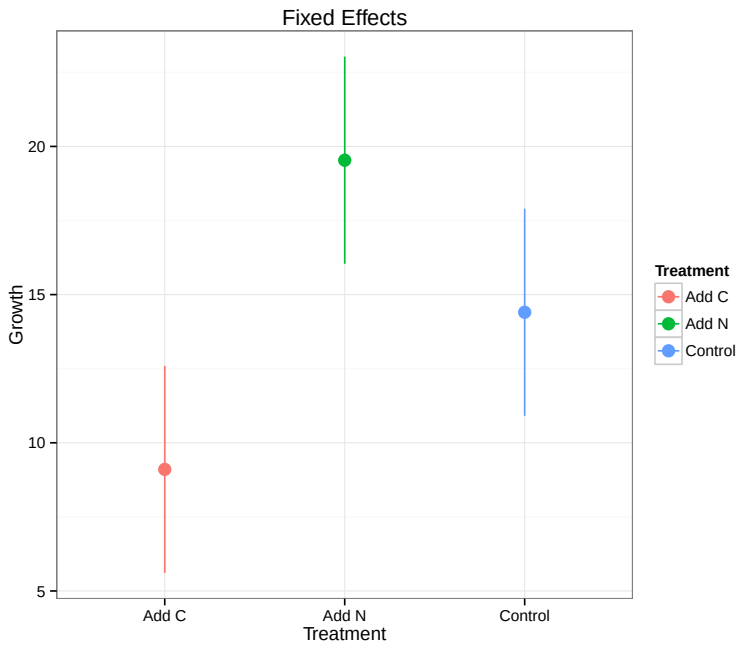
```
fixef(plantLME)

#      (Intercept)   TreatmentAdd N TreatmentControl
#         9.102532        10.432122         5.301297

coef(plantLME)

#   (Intercept) TreatmentAdd N TreatmentControl
# a    7.457340       10.43212         5.301297
# b   12.425751       10.43212         5.301297
# c   10.310714       10.43212         5.301297
# d    6.216325       10.43212         5.301297
# e    7.232907       10.43212         5.301297
# f   10.251140       10.43212         5.301297
# g   11.798648       10.43212         5.301297
# h    7.127435       10.43212         5.301297
# i   13.744547       10.43212         5.301297
# j    6.042597       10.43212         5.301297
# k   11.060393       10.43212         5.301297
```


BLUPs of Random Effects

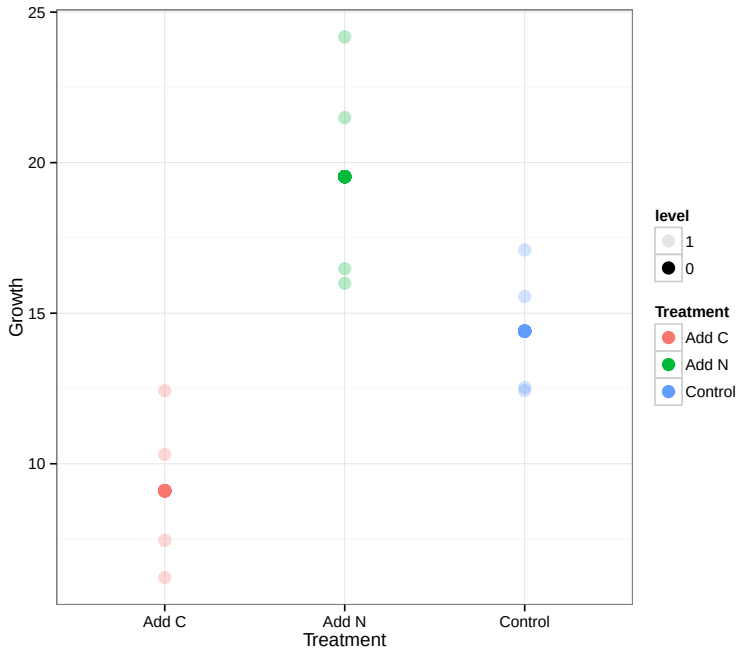
```
ranef(plantLME)

#   (Intercept)
# a   -1.645193
# b    3.323219
# c    1.208181
# d   -2.886207
# e   -1.869626
# f    1.148608
# g    2.696115
# h   -1.975097
# i    4.642015
# j   -3.059935
# k    1.957860
# l   -3.539941
```

Visualizing Fixed Effects

```
plantLME2 <- lme(Growth ~ Treatment-1, random = ~ 1|Pot, data=plants)
```



Visualizing Fixed and Random Effects



For more on confidence intervals, see
`http://glmm.wikidot.com/faq`

Visualizing Fixed and Random Effects

```
fixRanDF <- data.frame(Treatment = rep(plants$Treatment, 2),

                       level= factor(c(rep(0,36), rep(1, 36)),
                                     levels=c(1,0)),

                       Growth = c(fitted(plantLME, level=0),
                                  fitted(plantLME, level=1)))

qplot(Treatment, Growth, alpha=level, color=Treatment,
      data=fixRanDF, size=I(4)) +
    theme_bw()
```