Blocked designs, split plots, and intro. to linear mixed models

Malachy Campbell

Sept. 11, 2020

Recap

- **Simple linear regression:** $y = \beta_0 + \beta_1 x_1 + e$
- **Multiple linear regression:** $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + e$
- **One-way ANOVA:** $y = \mu + \alpha_1 x_{A1} + \alpha_2 x_{A2} + \ldots + e$
    - One factor - A; x_{A1} and x_{A2} are dummy variables (0,1) for the second and third level of factor A
- **Two-way ANOVA:** $y = \mu + \alpha_1 x_{A1} + \alpha_2 x_{A2} + \ldots + \beta_1 x_{B1} + \beta_2 x_{B2} + e$
    - Two factors - A,B; x_{B1} and x_{B2} are dummy variables (0,1) for the second and third levels of factor B

- x's are continuous in regression and binary (0,1) in ANOVA

Recap

- **Simple linear regression:** $y = \beta_0 + \beta_1 x_1 + e$
- **Multiple linear regression:** $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + e$
- **One-way ANOVA:** $y = \mu + \alpha_1 x_{A1} + \alpha_2 x_{A2} + \ldots + e$
    - One factor - A; x_{A1} and x_{A2} are dummy variables (0,1) for the second and third level of factor A
- **Two-way ANOVA:**
  $y = \mu + \alpha_1 x_{A1} + \alpha_2 x_{A2} + \ldots + \beta_1 x_{B1} + \beta_2 x_{B2} + e$
    - Two factors - A,B; x_{B1} and x_{B2} are dummy variables (0,1) for the second and third levels of factor B

Recap

$$\mathrm{SS}_{\mathrm{Total}} = \mathrm{SS}_{\mathrm{Regression}} + \mathrm{SS}_{\mathrm{Error}}$$

Does the model fit the data?

Recap

$$\mathrm{SS}_{\mathrm{Total}} = \mathrm{SS}_{\mathrm{Regression}} + \mathrm{SS}_{\mathrm{Error}}$$

Does the model fit the data?

Regression:

- H_0**:** All $\beta_i = 0$; $i = (1, 2, \ldots)$
- H_A**:** At least one $\beta_i \neq 0$; $i = (1, 2, \ldots)$

ANOVA:

- H_0**:** All $\beta_i = 0$; All $\alpha_i = 0$ $i = (1, 2, \ldots)$
- H_A**:** At least one β_i or $\alpha_i \neq 0$; $i = (1, 2, \ldots)$

Recap

$$\mathrm{SS_{Total}} = \mathrm{SS_{Regression}} + \mathrm{SS_{Error}}$$

Does the model fit the data?

Regression:

- H_0**:** All $\beta_i = 0$; $i = (1, 2, \ldots)$
- H_A**:** At least one $\beta_i \neq 0$; $i = (1, 2, \ldots)$

ANOVA:

- H_0**:** All $\beta_i = 0$; All $\alpha_i = 0$ $i = (1, 2, \ldots)$
- H_A**:** At least one β_i or $\alpha_i \neq 0$; $i = (1, 2, \ldots)$

$$F = \frac{SS_{reg.}/df_{reg.}}{SS_{err.}/df_{err.}} = \frac{MS_{reg.}}{MS_{err.}}$$

Compare to F distribution $(F(df_{reg.}, df_{err.}))$

Recap

$$\mathrm{SS}_{\mathrm{Total}} = \mathrm{SS}_{\mathrm{Regression}} + \mathrm{SS}_{\mathrm{Error}}$$

Does the model fit the data?

Regression:

- H_0: $y = \beta_0 + e$
- H_A: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i + e$

ANOVA:

- H_0: $y = \mu + e$
- H_A: $y = \mu + \alpha_1 x_{A1} + \alpha_2 x_{A2} + e$

Recap

$$\mathrm{SS}_{\mathrm{Total}} = \mathrm{SS}_{\mathrm{Regression}} + \mathrm{SS}_{\mathrm{Error}}$$

Does the model fit the data?

We are just comparing models

- H_0**:** Reduced model
- H_A**:** Full model

$$\mathrm{F} = \frac{(\mathrm{SSE(Reduced)} - \mathrm{SSE(Full)})/(df_{\mathrm{Reduced}} - df_{\mathrm{Full}})}{\mathrm{SSE(Full)}/df_{\mathrm{Full}}}$$

$$\mathrm{F}(df_{\mathrm{Reduced}} - df_{\mathrm{Full}}, df_{\mathrm{Full}})$$

Recap

Two-way ANOVA (Factors A and B) with no interaction:

$$\mathrm{SS}_{\mathrm{Total}} = \mathrm{SS}_{\mathrm{Regression}} + \mathrm{SS}_{\mathrm{Error}}$$
$$\mathrm{SS}_{\mathrm{Total}} = \mathrm{SS}_{\mathrm{A}} + \mathrm{SS}_{\mathrm{B}} + \mathrm{SS}_{\mathrm{Error}}$$

- We can only split the $\mathrm{SS}_{\mathrm{Regression}}$ into orthogonal components SS_{A} and SS_{B} if the design is **balanced** (all treatment combinations appear and equal number of times)
- When the design is unbalanced A and B are correlated, thus the effect of A cannot be completely separated from the effect of B

Recap

Two-way ANOVA (Factors A and B) with no interaction:

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$$

Type I:

- For factor A:
  - H_0: $y = \mu + e$ (Reduced)
  - H_1: $y = \mu + \alpha_i + e$ (Full)
- For Factor B:
  - H_0: $y = \mu + \alpha_i + e$ (Reduced)
  - H_1: $y = \mu + \alpha_i + \beta_j + e$ (Full)

$$\mathrm{F} = \frac{(\mathrm{SSE(Reduced)} - \mathrm{SSE(Full)})/(df_{\mathrm{Redu}}}{\mathrm{SSE(Full)}/df_{\mathrm{Full}}}$$

$$\mathrm{F}(df_{\mathrm{Reduced}} - df_{\mathrm{Full}}, df_{\mathrm{Full}})$$

Recap

Two-way ANOVA (Factors A and B) with no interaction:

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$$

Type II and III: Type II and III are equivalent since there are no interactions

- For factor A:
  - H_0: $y = \mu + \beta_j + e$ (Reduced)
  - H_1: $y = \mu + \alpha_i + \beta_j + e$ (Full)
- For Factor B:
  - H_0: $y = \mu + \alpha_i + e$ (Reduced)
  - H_1: $y = \mu + \alpha_i + \beta_j + e$ (Full)

Recap

Two-way ANOVA (Factors A and B) with with interaction:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

Recap

Two-way ANOVA (Factors A and B) with with interaction:

Type II:

- For factor A:
  - H_0: $y = \mu + \beta_j + e$ (Reduced)
  - H_1: $y = \mu + \alpha_i + \beta_j + e$ (Full)
- For Factor B:
  - H_0: $y = \mu + \alpha_i + e$ (Reduced)
  - H_1: $y = \mu + \alpha_i + \beta_j + e$ (Full)
- For Interaction:
  - H_0: $y = \mu + \alpha_i + \beta_j + e$ (Reduced)
  - H_1: $y = \mu + \alpha_i + \beta_j + \gamma_{ij} + e$ (Full)

Type III:

- For factor A:
  - H_0: $y = \mu + \beta_j + \gamma_{ij} + e$
  - H_1: $y = \mu + \alpha_i + \beta_j + \gamma_{ij} + e$
- For Factor B:
  - H_0: $y = \mu + \alpha_i + \gamma_{ij} + e$
  - H_1: $y = \mu + \alpha_i + \beta_j + \gamma_{ij} + e$
- For Interaction:
  - H_0: $y = \mu + \alpha_i + \beta_j + e$
  - H_1: $y = \mu + \alpha_i + \beta_j + \gamma_{ij} + e$

Recap

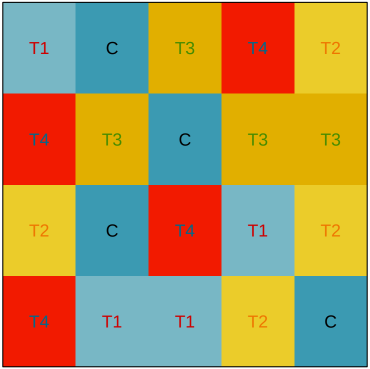
Two-way ANOVA (Factors A and B) with with interaction:

- For Type II, main effects for a given factor are tested in absence of interactions involving that factor

Blocking, split-plot designs and intro to mixed models

Completely randomized designs

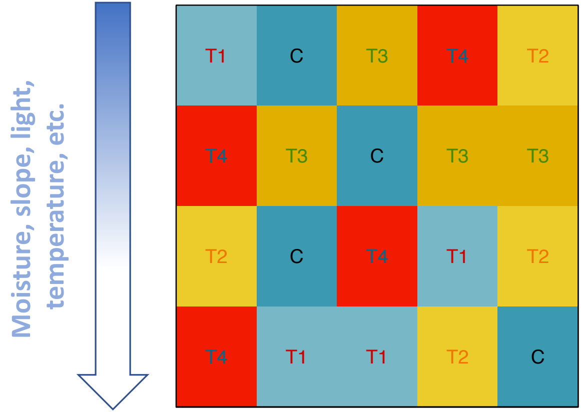
Levels of the factor are randomly assigned to experimental units



T1	C	T3	T4	T2
T4	T3	C	T3	T3
T2	C	T4	T1	T2
T4	T1	T1	T2	C

Completely randomized designs

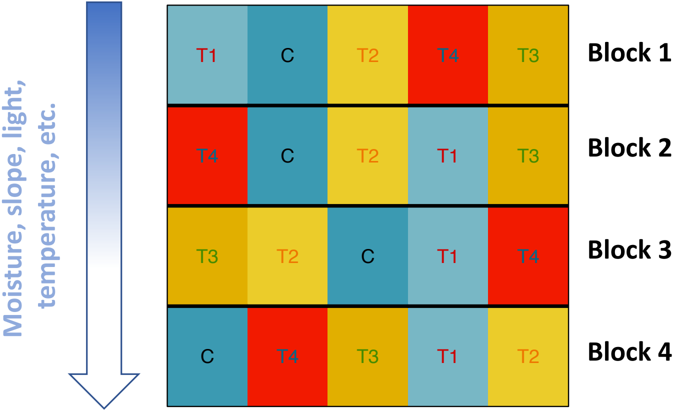
What if there is a known gradient in the field?



T1	C	T3	T4	T2
T4	T3	C	T3	T3
T2	C	T4	T1	T2
T4	T1	T1	T2	C

Completely randomized designs

Complete randomized block designs: Constrain randomization to control for unwanted variation.



Moisture, slope, light, temperature, etc.

T1	C	T2	T4	T3	**Block 1**
T4	C	T2	T1	T3	**Block 2**
T3	T2	C	T1	T4	**Block 3**
C	T4	T3	T1	T2	**Block 4**

Completely randomized block designs

Complete randomized block designs: Constrain randomization to control for unwanted variation.

$$T = No.treatment; \quad B = No.blocks$$

Source	**SS**	**df**	**MS**	**F-val**
Trt.	$SS_{Trt.}$	$df_{Trt.} = T - 1$	$MS_{Trt.} = \frac{SS_{Trt.}}{df_{Trt.}}$	$F_{Trt.} = \frac{MS_{Trt.}}{MS_{Err.}}$
Blk.	$SS_{Blk.}$	$df_{Blk.} = B - 1$	$MS_{Blk.} = \frac{SS_{Blk.}}{df_{Blk.}}$	$F_{Blk.} = \frac{MS_{Blk.}}{MS_{Err.}}$
Err.	$SS_{Err.}$	$df_{Err.} = (T-1)(B-1)$	$MS_{Err.} = \frac{SS_{Err.}}{df_{Err.}}$	
Total	$SS_{Tot.}$	$df_{Tot.} = (T \times B - 1)$		

$$df_{Err.} = (T-1)(B-1) = n - (T-1) - (B-1) - 1$$

Completely randomized block designs

Complete randomized block designs: Constrain randomization to control for unwanted variation.

$$T = No.treatment; \quad B = No.blocks$$

Source	SS	df	MS	F-val
Trt.	$SS_{Trt.}$	$df_{Trt.} = T - 1$	$MS_{Trt.} = \frac{SS_{Trt.}}{df_{Trt.}}$	$F_{Trt.} = \frac{MS_{Trt.}}{MS_{Err.}}$
Blk.	$SS_{Blk.}$	$df_{Blk.} = B - 1$	$MS_{Blk.} = \frac{SS_{Blk.}}{df_{Blk.}}$	$F_{Blk.} = \frac{MS_{Blk.}}{MS_{Err.}}$
Err.	$SS_{Err.}$	$df_{Err.} = (T-1)(B-1)$	$MS_{Err.} = \frac{SS_{Err.}}{df_{Err.}}$	
Total	$SS_{Tot.}$	$df_{Tot.} = (T \times B - 1)$		

$$df_{Err.} = (T-1)(B-1) = n - (T-1) - (B-1) - 1$$

- Is blocking necessary? → Check $F_{Blk.}$
  - If not, we are wasting DF estimating block effects = lower power for treatment effects.

Completely randomized block designs

Complete randomized block designs: Constrain randomization to control for unwanted variation.

$$T = No.treatment; \quad B = No.blocks$$

Source	**SS**	**df**	**MS**	**F-val**
Trt.	$SS_{Trt.}$	$df_{Trt.} = T - 1$	$MS_{Trt.} = \frac{SS_{Trt.}}{df_{Trt.}}$	$F_{Trt.} = \frac{MS_{Trt.}}{MS_{Err.}}$
Blk.	$SS_{Blk.}$	$df_{Blk.} = B - 1$	$MS_{Blk.} = \frac{SS_{Blk.}}{df_{Blk.}}$	$F_{Blk.} = \frac{MS_{Blk.}}{MS_{Err.}}$
Err.	$SS_{Err.}$	$df_{Err.} = (T - 1)(B - 1)$	$MS_{Err.} = \frac{SS_{Err.}}{df_{Err.}}$	
Total	$SS_{Tot.}$	$df_{Tot.} = (T \times B - 1)$		

$$df_{Err.} = n - (T - 1) - (B - 1) - 1$$

$$df_{Err.} = \text{No. obs.} - \text{Trt. df.} - \text{Blk. df.} - 1$$

Interaction DF: $(B - 1)(T - 1)$

- Can we check for the interaction between block and treatment?

Completely randomized block designs

Complete randomized block designs: Constrain randomization to control for unwanted variation.

$$T = No.treatment; \quad B = No.blocks$$

Source	SS	df	MS	F-val
Trt.	$SS_{Trt.}$	$df_{Trt.} = T - 1$	$MS_{Trt.} = \frac{SS_{Trt.}}{df_{Trt.}}$	$F_{Trt.} = \frac{MS_{Trt.}}{MS_{Err.}}$
Blk.	$SS_{Blk.}$	$df_{Blk.} = B - 1$	$MS_{Blk.} = \frac{SS_{Blk.}}{df_{Blk.}}$	$F_{Blk.} = \frac{MS_{Blk.}}{MS_{Err.}}$
Err.	$SS_{Err.}$	$df_{Err.} = (T-1)(B-1)$	$MS_{Err.} = \frac{SS_{Err.}}{df_{Err.}}$	
Total	$SS_{Tot.}$	$df_{Tot.} = (T \times B - 1)$		

$$df_{Err.} = n - (T-1) - (B-1) - 1$$

$$df_{Err.} = \text{No. obs.} - \text{Trt. df.} - \text{Blk. df.} - 1$$

Interaction DF: $(B-1)(T-1)$

- Can we check for the interaction between block and treatment? → Only if we have replication in blocks

Completely randomized block designs

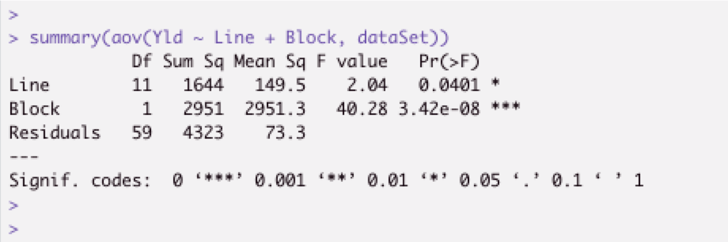
- Analysis of RCBD can be done using basic two-way ANOVA
  - If unbalanced, then use appropriate ANOVA (Type II or Type III)
- Blocks that are too big may not be optimal
- If there are a large number of blocks, then an ANOVA may not be the best approach (we are using many DF to estimate effects for a "nuicance" factor).

Analyzing a RCBD in R

We are interested in evaluating yield for 12 wheat varieties. The experiment is laid out as a complete randomized block design with six blocks. All 12 varieties were grown in each block. The design is balanced.

This is just a straight-forward two-way ANOVA.

- `summary(aov(Yld ~ Line + Block, dataSet))`



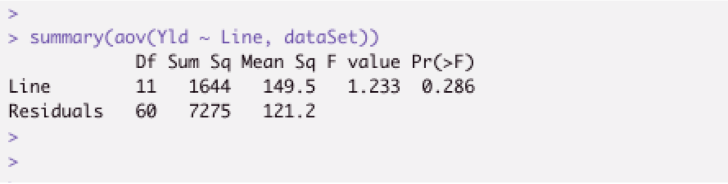
```
>
> summary(aov(Yld ~ Line + Block, dataSet))
            Df Sum Sq Mean Sq F value   Pr(>F)    
Line        11   1644   149.5    2.04   0.0401 *  
Block        1   2951  2951.3   40.28 3.42e-08 ***
Residuals   59   4323    73.3                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
>
```

Analyzing a RCBD in R

We are interested in evaluating yield for 12 wheat varieties. The experiment is laid out as a complete randomized block design with six blocks. All 12 varieties were grown in each block. The design is balanced.

What happens if we treat this as a completely randomized design?

- `summary(aov(Yld ˜ Line, dataSet))`



```
>
> summary(aov(Yld ~ Line, dataSet))
            Df Sum Sq Mean Sq F value Pr(>F)
Line        11   1644   149.5   1.233  0.286
Residuals   60   7275   121.2
>
>
```

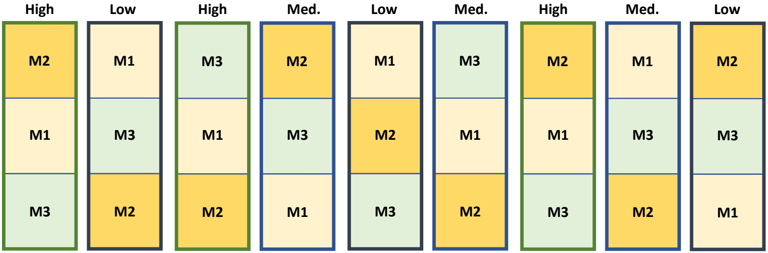
Split-plot designs

Say we are interested in investigating the effects of drought on the growth of three mutant lines (M1, M2, M3). We have three levels irrigation levels: low (25% field capacity), medium (40% field capacity), and high (65% field capacity). We are planning to record three observations for each irrigation and line combination. Thus, we will have $(3 \times 3 \times 3 = 27)$ plots (experimental units) with 20 plants in each plot.

How can we apply irrigation to a small plot? Can we ensure that the treatment doesn't influence neighboring plots?

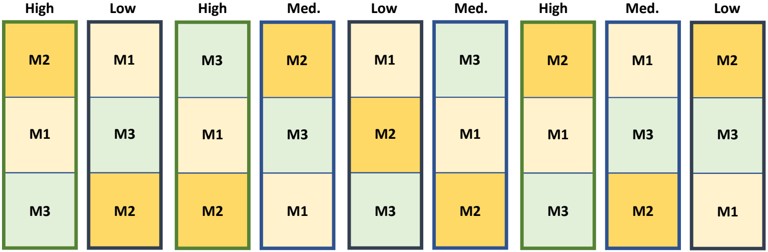
Split-plot designs

Say we are interested in investigating the effects of drought on the growth of three mutant lines (M1, M2, M3). We have three levels irrigation levels: low (25% field capacity), medium (40% field capacity), and high (65% field capacity). We are planning to record three observations for each irrigation and line combination. Thus, we will have ($3 \times 3 \times 3 = 27$) plots (experimental units) with 20 plants in each plot.



High	Low	High	Med.	Low	Med.	High	Med.	Low
M2	M1	M3	M2	M1	M3	M2	M1	M2
M1	M3	M1	M3	M2	M1	M1	M3	M3
M3	M2	M2	M1	M3	M2	M3	M2	M1

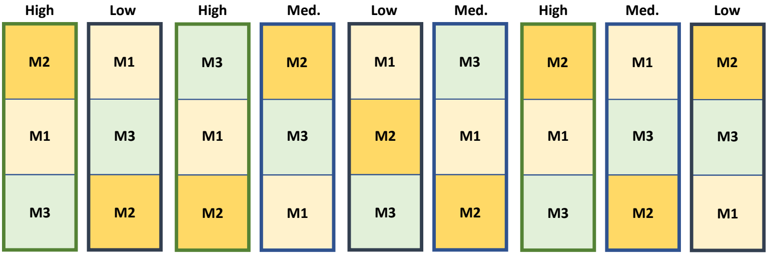
Split-plot designs



High	Low	High	Med.	Low	Med.	High	Med.	Low
M2	M1	M3	M2	M1	M3	M2	M1	M2
M1	M3	M1	M3	M2	M1	M1	M3	M3
M3	M2	M2	M1	M3	M2	M3	M2	M1

- Two (independent) randomizations
  - Irrigation is randomly assigned to **whole plots**
  - Lines are randomly assigned **subplots** within whole plots

Split-plot designs



High	Low	High	Med.	Low	Med.	High	Med.	Low
M2	M1	M3	M2	M1	M3	M2	M1	M2
M1	M3	M1	M3	M2	M1	M1	M3	M3
M3	M2	M2	M1	M3	M2	M3	M2	M1

- Two (independent) randomizations
  - Irrigation is randomly assigned to **whole plots**
  - Lines are randomly assigned **subplots** within whole plots
- One way to think of this design is two experiments in one
  - Completely randomized design for irrigation
    - Whole plots are experimental units for irrigation factor
  - Sub-plots are experimental units for the factor mutant

Split plot designs

Two experiments in one:

- Completely randomized design for irrigation (whole plot are experimental units)
    - **Remember in a CRD we assume error for exp. units is** $e \sim N(0, \sigma^2)$
    - We will have a random error for the whole plots!
- Sub-plots are experimental units for the factor mutant
    - We will also have a random error for the sub-plots!

Split plot designs

What would this model look like?

- So far we've only encountered models with one random term on the right hand side (error term; $e \sim N(0, \sigma^2)$)
- Mixed linear models (mixed-effects models) allow us to fit more than one random effect

Split plot designs

What would this model look like?

The model:

$$y_{ijk} = \mu + \alpha_i + u_k + \beta_j + \gamma_{ij} + e_{k(ij)}$$

Random Terms:

- y_{ijk}: Growth for ith irrigation level, jth mutant in kth main plot
- u_k: Random effect for the kth main plot $u_k \sim N(0, \sigma_u^2)$
- $e_{k(ij)}$: Random error for subplot $e_{k(ij)} \sim N(0, \sigma^2)$

Fixed Terms:

- α_i: Effect for the ith irrigation level
- β_j: Effect for the jth mutant line
- γ_{ij}: Interaction effect between the ith irrigation level and jth mutant line

Split plot designs

What would this model look like?

The model:

$$y_{ijk} = \mu + \alpha_i + u_k + \beta_j + \gamma_{ij} + e_{k(ij)}$$

In plain English:
The observations for each combination and irrigation level will randomly from their mean (determined by the fixed effects in the model). This random deviation is due to:

1. Whole plot error u_k
   - Note that all observations in the same whole plot share the same random effect (i.e. u_k)
2. Sub-plot error $e_{k(ij)}$

Split plot designs

What would this model look like?

The model:

$$y_{ijk} = \mu + \alpha_i + u_k + \beta_j + \gamma_{ij} + e_{k(ij)}$$

The group means (fitted values;) for a model with all fixed effects (OLS) are called **B**est **L**inear **U**nbiased **E**stimates (BLUEs). The fitted values obtained with a mixed model are **B**est **L**inear **U**nbiased **P**redictions (BLUPs) because they are calculated using a random effect (u_i).

Mixed linear models

Mixed models are extremely useful and their utility extends far beyond split plot designs.

- Blocking
- Repeated measures
    - Subsamples, technical replicates from the same experimental unit
- Longitudinal data
    - Measurements on the same subjects over time
- Many, many more

Mixed linear models

Mixed models are extremely useful and their utility extends far beyond split plot designs.

Should this term be fixed or random?

- Are you interested in making inferences about the levels of the factor? Yes → fixed
- Are you interested in making inferences about the population in which the factor was drawn? Yes → random
- How many levels of the factor are there? Do you have enough degrees of freedom? Is it worth it?
    - Are you interested the estimating effects for 20 blocks? No → random

Fitting Linear Mixed Models

Fitting a mixed model is less straight forward than ordinary least squares. Model parameters are estimated using numerical optimization.

- Numerical optimization: Iterative approach to maximize the likelihood function for the model (PDF of the the data given the parameters)
  - Choose some values for the parameter that maximize the likelihood of observing the data
- **Maximum likelihood:** Estimates for variance components are biased
  - Unknown estimate of the mean is used to compute estimate for variance components
- **Restricted (or Residual) Maximum likelihood:** Variance components are unbiased
  - Optimize a log likelihood function that does not depend on the mean

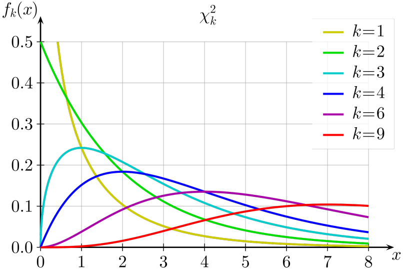
Hypothesis tests for Linear Mixed Models

We can test for fixed effects using:

- Likelihood ratio test (LRT): Model comparisons
  - Fit "full" model → drop one fixed term → compare log likelihoods between full and reduced

    $$\lambda_{LR} = -2[l(\hat{\theta})_{Red.} - l(\hat{\theta})_{Full}]$$

  - λ_{LR} follows a chi-squared (χ^2) distribution with k = no. fixed terms that differ between models.



Hypothesis tests for Linear Mixed Models

We can test for fixed effects using:

- Wald's test: Similar to a Z or t-test
  - Scale the effect by the standard error of the effect (this should sound familiar)

$$t = \frac{(\hat{\theta} - \theta_o)^2}{\hat{\mathrm{SE}}}$$

Assessing goodness of fit

Akaike information criterion:

- Log likelihood "penalized" by model complexity ($k =$ no. parameters in model)

$$\mathrm{AIC} = 2k - 2\mathrm{logLik}$$

Bayesian information criterion:

- Similar to AIC, but with a heavier penalty ($klog(n)$; n is the number of observations)

$$\mathrm{AIC} = k\ln(n) - 2\mathrm{logLik}$$

Lower values indicate a better fit.