Training course-cum-workshop on

Statistical, Biometrical and Genomic Methods

Mexico city, May 21, 2019

Alvarado Beltran Gregorio: g.alvarado@cigar.org

CIMMYT
International Maize and Wheat Improvement Center

Fixed Effects Models

- In the linear models for the basic experimental designs (CRD, for example)
- $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$

- The parameters of the treatments effects have been treated as fixed, These models are called *fixed-effects* models, because the treatment effects are fixed numbers

- **But sometimes the fixed-effects assumptions don't make much sense , especially when we need to be able to make inferences for the whole population, not just the random sample that we used in the experiment.**

Random Effects

Random effects **are another approach to designing experiments and modeling data.**

Random effects are appropriate when the treatments are random samples from a population of potential treatments.

They are also useful for random subsampling from populations.

Random-effects models assume that the treatment effects are random variables.

Also, the focus of inference is on the population, not the individual treatment effects

Random Effects Model

The basic random effects model begins with the usual decomposition:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

Now we are assuming that both treatment effects and residuals are random

Again we assume that the ε_{ij} are independent and normally distributed with mean 0 and variance σ^2, as we did in fixed effects,

We also assume that the treatment effects τ_i are independent and normally distributed with mean 0 and variance ${\sigma_\tau}^2$, and that the τ_i's and the ε_{ij}'s are independent of each other.

Random Effects variance components

- So the variance of y_{ij} is $\sigma_\tau{}^2 + \sigma^2$. The terms $\sigma_\tau{}^2$ y σ^2 are called *components of variance* or *variance components*. Thus the random-effects model is sometimes called a *components of variance model*
- The parameters of the random effects model are the overall mean μ, the error variance σ^2, and the variance of the treatment effects $\sigma_\tau{}^2$; but now the treatment effects τ_i are random variables, not fixed parameters.
- We want to make inferences about these parameters; Typical inferences would be point estimates or confidence intervals for the variance components, or a test of the null hypothesis that the treatment variance $\sigma_\tau{}^2 = 0$.

Selecting Fixed or Random Effects

- **We must decide whether each factor is fixed or random. This decision is usually straightforward but can actually vary depending upon the goals of an experiment.**

- **Suppose that we have an animal breeding experiment with four sires. Now we know that the four sires we used are the four sires that were available; we did no random sampling from a population.**

- **If we are trying to make inferences about just these four sires, we treat sire as a fixed effect. On the other hand, if we are trying to make inferences about the population of potential sires, we would treat sires as a random effect.**

Mixed Effects Model

An experiment with both fixed and random effects is said to have *mixed* effects.

The interaction of a fixed effect and a random effect must be random, because a new random sample of factor levels will also lead to a new sample of interactions.

This model must reflect both the structure of the experiment (nesting and/or crossing of effects), how broadly we are trying to make inference (just these treatments or a whole population of treatments)

Once we have answered these questions, we can build a model. Parameters are only defined within a model, so we need the model to make tests, compute confidence intervals, and so on.

- A simple model for a continuous trait using a RCBD is:

$$y_{ij} = \mu + G_i + R_j + \varepsilon_{ij}$$

- where y_{ij} is the trait value for genotype i in replication j, μ is the population mean, G_i is the genetic effect for genotype i , R_j is the effect of the replication j, and ε_{ij} is the error term associated with genotype i in replication j.

- For a RIBD or Lattice, the model becomes:

$$y_{ijk} = \mu + G_i + R_j + B_k\,(R_j) + \varepsilon_{ijk}$$

- where now the additional term $B_k\,(R_j)$, is the effect of the incomplete block k, nested within the replication j

- If it is assumed that the components in the model are distributed as normal variables
- $$y \sim N(\mu, \sigma^2_p)$$
$$G \sim N(\mu, \sigma^2_g)$$
$$\varepsilon \sim N(0, \sigma^2_e)$$
- and the covariance between genetic effects and experimental error is zero, then

$$\sigma^2_p = \sigma^2_g + \sigma^2_e$$

- If the same genotype is replicated r times in an experiment and phenotypic means are used, then the relationship become
- $$\sigma^2_p = \sigma^2_g + \sigma^2_e / r$$

- If the same genotypes are tested in several environments, such as locations or years, then the simple model for RCBD can be extended to:

$$y_{ijk} = \mu + G_i + E_j + (GE)_{ij} + R_k(E_j) + \varepsilon_{ijk}$$

- where E_j and $(GE)_{ij}$ are the environmental effects and genotype by environmental interactions.

- Similarly for RIBD the model becomes:

- $y_{ijkl} = \mu + G_i + E_j + (GE)_{ij} + R_k(E_j) + B_l(R_k E_j) + \varepsilon_{ijkl}$

How calculate BLUEs and/or BLUPs?

Many statistical software packages can compute expected mean squares for unbalanced data, but most do not compute all the possibilities. For example, SAS PROC GLM can compute Type I, II, or III expected mean squares

When we have balanced data without missing values and we are interested in calculate the Best Linear Unbiased Estimators (BLUE's) or adjusted means, we must declare the factor of interest as Fixed Effect, in both Fixed Effects or Mixed Effects Models

Example

RESEARCH PROGRAM ON Maize
CGIAR

Results Means and LSMeans as RCB

Level	Mean	gen	yieldLSMe	GenMixec	Estimate
1	5.1625	1	5.1625	1	5.1625
5	5.0644	5	5.0644	5	5.0644
12	4.9144	12	4.9144	12	4.9144
15	4.8933	15	4.8933	15	4.8933
19	4.8656	19	4.8656	19	4.8656
13	4.8294	13	4.8294	13	4.8294
21	4.8157	21	4.8157	21	4.8157
17	4.7285	17	4.7285	17	4.7285
16	4.7261	16	4.7261	16	4.7261
6	4.7103	6	4.7103	6	4.7103
22	4.6383	22	4.6383	22	4.6383
14	4.5584	14	4.5584	14	4.5584
2	4.5142	2	4.5142	2	4.5142
18	4.4404	18	4.4404	18	4.4404
4	4.4007	4	4.4007	4	4.4007
10	4.3889	10	4.3889	10	4.3889
11	4.3847	11	4.3847	11	4.3847
8	4.3245	8	4.3245	8	4.3245
24	4.1424	24	4.1424	24	4.1424
23	4.1362	23	4.1362	23	4.1362
7	4.1299	7	4.1299	7	4.1299
20	3.7849	20	3.7849	20	3.7849
9	3.6116	9	3.6116	9	3.6116
3	3.3431	3	3.3431	3	3.3431

CIMMYT
International Maize and Wheat Improvement Center

RESEARCH PROGRAM ON Maize
CGIAR

Results Means and LSMeans as Lattice

Level	Mean	gen	yieldLSMe	GenMixed	Estimate
1	5.1625	1	5.0760	1	5.1077
5	5.0644	5	5.0329	5	5.0372
12	4.9144	15	5.0154	15	4.9691
15	4.8933	14	4.9039	19	4.8403
19	4.8656	19	4.8440	21	4.7950
13	4.8294	21	4.7610	14	4.7757
21	4.8157	13	4.7329	13	4.7579
17	4.7285	16	4.7232	12	4.7553
16	4.7261	8	4.6652	16	4.7301
6	4.7103	12	4.6427	17	4.6026
22	4.6383	4	4.5354	6	4.5367
14	4.5584	17	4.5107	8	4.5276
2	4.5142	2	4.4726	22	4.5275
18	4.4404	22	4.4596	4	4.4901
4	4.4007	6	4.4255	2	4.4785
10	4.3889	10	4.3596	10	4.3732
11	4.3847	18	4.3173	18	4.3617
8	4.3245	23	4.3135	11	4.2833
24	4.1424	11	4.2184	23	4.2524
23	4.1362	20	4.1975	24	4.1539
7	4.1299	24	4.1396	7	4.1111
20	3.7849	7	4.1107	20	4.0400
9	3.6116	3	3.6110	9	3.5022
3	3.3431	9	3.4398	3	3.4992

CIMMYT
International Maize and Wheat Improvement Center

When we use BLUPs?

- **When data are unbalanced is highly recommended to use the Best Linear Unbiased Predictors (BLUPs) instead of the BLUEs, declaring the factor of interest as Random Effect in a Mixed Model**

- **Also when we are interested in calculate the genotypic variance component for broad sense heritability (repeatability), genetic correlations, etc., it necessary to declare the genotype effect as random effect**

RESEARCH PROGRAM ON Maize
CGIAR

BLUEs & BLUPs Comparisons

Gen	BLUEs	Grand_Me	REffect	BLUP
1	5.108	4.480	0.501	4.981
2	4.479	4.480	0.005	4.484
3	3.499	4.480	-0.785	3.695
4	4.490	4.480	0.006	4.486
5	5.037	4.480	0.475	4.954
6	4.537	4.480	0.045	4.524
7	4.111	4.480	-0.309	4.171
8	4.528	4.480	0.062	4.542
9	3.502	4.480	-0.810	3.670
10	4.373	4.480	-0.089	4.390
11	4.283	4.480	-0.196	4.283
12	4.755	4.480	0.226	4.705
13	4.758	4.480	0.232	4.711
14	4.776	4.480	0.243	4.723
15	4.969	4.480	0.425	4.904
16	4.730	4.480	0.201	4.680
17	4.603	4.480	0.078	4.558
18	4.362	4.480	-0.110	4.369
19	4.840	4.480	0.290	4.769
20	4.040	4.480	-0.339	4.141
21	4.795	4.480	0.256	4.736
22	4.528	4.480	0.024	4.504
23	4.252	4.480	-0.177	4.303
24	4.154	4.480	-0.253	4.226

CIMMYT
International Maize and Wheat Improvement Center

HERITABILITY

- Heritability in a broad sense is defined as the ratio of genotypic to phenotypic variance

$$H = \frac{\sigma^2_g}{\sigma^2_p} = \frac{\sigma^2_g}{\sigma^2_g + \sigma^2_e}$$

- Narrow sense heritability is defined as the ratio of the additive portion of genetic variance to the phenotypic variance

$$H = \frac{\sigma^2_a}{\sigma^2_p} = \frac{\sigma^2_a}{\sigma^2_a + \sigma^2_d + \sigma^2_I + \sigma^2_e}$$

- If phenotypic means are used, then we can obtain the mean based heritability

$$H = \frac{\sigma^2_g}{\sigma^2_p} = \frac{\sigma^2_g}{\sigma^2_g + \frac{1}{r}\sigma^2_e}$$

$$H = \frac{\sigma^2_g}{\sigma^2_p} = \frac{\sigma^2_g}{\sigma^2_g + \frac{1}{(e)}\sigma^2_{ge} + \frac{1}{(e)(r)}\sigma^2_e}$$

- **Genetic Correlations:** For two related traits, the values can be modelled as

- $y_{1j} = \mu_1 + G_{1i} + \varepsilon_{1j}$
- $y_{2j} = \mu_2 + G_{2i} + \varepsilon_{2j}$

- To quantify relationship between the two traits, matrices

$$\Sigma_p = \begin{bmatrix} \sigma^2_{p1} & \sigma^2_{p12} \\ \sigma^2_{p12} & \sigma^2_{p2} \end{bmatrix} = \Sigma_g + \Sigma_e = \begin{bmatrix} \sigma^2_{g1} & \sigma^2_{g12} \\ \sigma^2_{g12} & \sigma^2_{g2} \end{bmatrix} + \begin{bmatrix} \sigma^2_{e1} & \sigma^2_{e12} \\ \sigma^2_{e12} & \sigma^2_{e2} \end{bmatrix}$$

- are used, where Σ_p, Σ_g and Σ_e are variance-covariance matrices for phenotypic, genetic and environmental effects, respectively

- The relationship between the two traits can be quantified by

$$\rho_p = \frac{\sigma_{p12}}{\sqrt{\sigma^2_{p1}\sigma^2_{p2}}}$$

$$\rho_g = \frac{\sigma_{g12}}{\sqrt{\sigma^2_{g1}\sigma^2_{g2}}}$$

$$\rho_e = \frac{\sigma_{e12}}{\sqrt{\sigma^2_{e1}\sigma^2_{e2}}}$$

- where ρ_p , ρ_g and ρ_e are defined as phenotypic, genetic and environmental correlations, respectively

• *Sum Method of Estimating Covariance*

- The Sum Method is based on the statistical property of the sum of two random variables, which states:

$$Var\ (X+Y) = Var\ (X) + Var\ (Y) + 2\ Cov\ (X,Y)$$

- This can be rearranged and written as

$$Cov\ (X,Y)=[Var\ (X+Y) - Var\ (X) - Var\ (Y)\ ] / 2$$

- In this case X and Y refer to two different traits evaluated within the same population.

- Using PROC MIXED in SAS and LME4 in R we can get an estimate for the variance components for Var (X), Var (Y) and Var (X+Y).

- The Var (X+Y) is obtained by running MIXED on a new variable created by summing the plot mean values for trait X and trait Y

- Genetic correlations among sites
  - Cooper, DeLacy & Basford (1996)

$$\sigma^2_{ge} = \frac{Suma_{jj'}\{[\sigma^2_{g(j)} - \sigma^2_{g(j')}]^2 + 2\sigma^2_{g(j)}\sigma^2_{g(j')}[1-\rho_{g(jj')}]\}}{(e)(e-1)}$$

$$\rho_{p(jj')} = (h_j)(h_{j'})\rho_{g(jj')}$$

- META-R (Multi-Environment Trial Analysis using R)
  - is a set of codes developed in the statistical software R
  - It is able to perform three kind of analysis
    - Estimate BLUEs and BLUPs as well some genetic parameters as heritability and variance components,
    - Compute Genetic correlations among locations,
    - Compute Genetic correlations between traits.


META-R (Multi Environment Trial Analysis with R for Windows)
Back
Analyze
Help
Analyses Options
none
none
Genetic Correlations among Locations
Genetic Correlations among Variables
BLUEs and BLUPs
Response Variables
Select one or more
Selected
Management
Plot
gyf
ad
asi
ph
eh
epo
gyf
ad
asi
ph
eh
epo
Select All
Drop All
Welcome to META-R (Multi Environment Trial Analysis with R for Windows). Version 6.0 (2016-11-30)
Copyright © 2016 Centro Internacional de Mejoramiento de Maíz y Trigo (CIMMYT).
Authors:
Gregorio Alvarado
Marco López
Mateo Vargas
Angela Pacheco
Francisco Rodríguez
Juan Burgueño
José Crossa
This program is based in some components from Java, developed by ORACLE AMERICA, INC. and R, developed by Kurt Hornik. Any Java component of this program is hereby licensed under the Oracle Binary Code License Agreement for the Java SE Platform Products and JavaFX made available by Oracle (available at http://www.oracle.com/technetwork/java/javase/terms/license/index.html). Any R component of this program as well as the program as a whole developed by CIMMYT are hereby licensed as per the terms of the GNU General Public License version 3(available at http://www.gnu.org/licenses/gpl-3.0.html), as specified below.
This program is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 3 of the License, or any later.
This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU

- Also it is possible to include some covariate(s), to correct the response, it could be number of plants or precocity (days to flowering), for example

- Linear models included in META-R are,

One site

Random Complete Block Designs without covariance analysis

- $Y_{ij} = \mu + \text{Rep}_j + \text{Gen}_i + \varepsilon_{ij}$

Random Complete Block Designs with covariance analysis

- $Y_{ij} = \mu + \text{Rep}_j + \text{Gen}_i + \text{Cov} + \varepsilon_{ij}$