Lecture 11

Random and Mixed Effects Models, Computer Experiments

Reading: Oehlert 2010 Chapter 11.1-11.7; DAE 2017 Chapters 17.1-17.3; 17.7& 20

DSA 8020 Statistical Methods II

Whitney Huang
Clemson University

Agenda

1. **Random and Mixed Effects Models**

2. **Computer Experiments**

Fixed Effects

Most settings we have dealt with so far (except the analysis of split-plot designs) have involved fixed effects:

CRD:	$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$
RCBD:	$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$
Factorial:	$y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ij}$

- The treatment effects are unknown but constants $\Rightarrow$ if we ran the experiment over again, would expect the same treatment effects
- We can increase the power of all of our tests by increasing the sample size n
- We perform inference on the treatment effects via t-tests and F-tests

Random Effects

Random effects models look very similar to fixed effects models. For example, we could have

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}.$$

The difference is in the assumptions we make for the treatment effects

Fixed Effects

Treatment effects $\alpha_i s$ are unknown constants that add to zero (or some other constraint)

Random Effects

- $\alpha_i s \sim \mathrm{N}(0, \sigma_\alpha^2)$
- $\alpha_i s$ are independent of ϵ_{ij}

How and Why Are Things So Different?

Fixed effects:

- The treatments are the treatments and they are unchanging
- If we rerun the experiment, we are still studying the same treatments

Random effects:

- The treatments are a random sample from a population of potential treatments
- If we rerun the experiment, we are looking at an entirely new sample of treatments
- Inference is on the population of potential treatments

Variance Components

Fixed effects:

- $\mathrm{Var}(y_{ij}) = \sigma^2$
- All $y_{ij}s$ are independent of each other
- Interest is about $\alpha_i s$

Random effects:

- $\mathrm{Var}(y_{ij}) = \sigma_\alpha^2 + \sigma^2$
- $\mathrm{Cor}(y_{ij}, y_{kl}) = \begin{cases} 0 & \text{if } i \neq k \\ \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma^2} & \text{if } i = k; j \neq l \\ 1 & \text{if } i = k; j = l \end{cases}$
- Interest is (mostly) about σ_α^2

An Example of Fixed Effects vs Random Effects

1. Compare reading ability of 10 2nd grade classes in NY:

   Select $g = 10$ specific classes of interest. Randomly choose n students from each classroom. Want to compare $\alpha_i s$ (class-specific effects) ⇒ Fixed effects

2. Compare variability among all 2nd grade classes in NY:

   Randomly choose $g = 10$ classes from large number of classes. Randomly choose n students from each classroom. Want to assess σ_α^2 (class to class variability) ⇒ Random effects

Random Effects Model (CRD)

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

where

- μ is the overall mean
- α_i: ith treatment effect and $\alpha_i \sim \mathrm{N}(0, \sigma^2_\alpha)$
- $\{\alpha_i\}$ and $\{\epsilon_{ij}\}$ independent
- The hypotheses are:

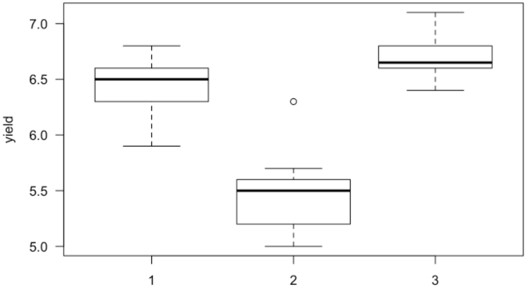
$$H_0 : \sigma^2_\alpha = 0$$
$$H_a : \sigma^2_\alpha > 0$$

One can use either "old school" method (ANOVA) or "new school" method (REML) to make inference about σ^2_α

Random Effects Example

Suppose that an agronomist is studying a large number of varieties of soybeans for yield. The agronomist randomly selects three varieties, and then randomly assigns each of those varieties to 10 of the 30 available plots

Soybean	Yield
V1	6.6, 6.4, 5.9, 6.6, 6.2, 6.7, 6.3, 6.5, 6.5, 6.8
V2	5.6, 5.2, 5.3, 5.1, 5.7, 5.6, 5.6, 6.3, 5.0, 5.4
V3	6.9, 7.1, 6.4, 6.7, 6.5, 6.6, 6.6, 6.6, 6.8, 6.8



Fixed Effects Analysis


```
> fixef <- lm(yield ~ var)
> anova(fixef)
Analysis of Variance Table

Response: yield
          Df Sum Sq Mean Sq F value    Pr(>F)
var        2  8.306  4.1530  49.593 9.114e-10 ***
Residuals 27  2.261  0.0837
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> coefficients(fixef)
(Intercept)        var2        var3
       6.45       -0.97        0.25
```

Random Effects Analysis

```
> library(lme4)
> randef <- lmer(yield ~ 1 + (1|var), REML = TRUE)
> summary(mod1)
Linear mixed model fit by maximum likelihood . t-tests
  use Satterthwaite's method [lmerModLmerTest]
Formula: yield ~ 1 + (1 | var)

     AIC      BIC   logLik deviance df.resid
    27.2     31.4    -10.6     21.2       27

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.8755 -0.6033  0.1245  0.5068  2.7574

Random effects:
 Groups   Name        Variance Std.Dev.
 var      (Intercept) 0.26849  0.5182
 Residual             0.08374  0.2894
Number of obs: 30, groups:  var, 3

Fixed effects:
            Estimate Std. Error     df t value Pr(>|t|)
(Intercept)   6.2100     0.3038 3.0000   20.44 0.000256
```

Concrete Cylinder Example Revisited

Suppose you are manufacturing concrete cylinders for bridge supports. There are three ways of drying concrete (say A, B, and C), and you want to find the one that gives you the best compressive strength. The concrete is mixed in batches that are large enough to produce exactly three cylinders, and your production engineer believes that there is substantial variation in the quality of the concrete from batch to batch.

	Batch					
Treatment	1	2	3	4	5	Trt Sum
A	52	47	44	51	42	236
B	60	55	49	52	43	259
C	56	48	45	44	38	231
Batch Mean	168	150	138	147	123	726

If we were treat the batch effects as random effects, then we have a Mixed Effects Model

Concrete Cylinder Example: Mixed Effects Analysis

```
> randef <- lmer(x ~ trt + (1|blk), REML = TRUE, data = dat)
> summary(randef)
Linear mixed model fit by REML. t-tests use
  Satterthwaite's method [lmerModLmerTest]
Formula: x ~ trt + (1 | blk)
   Data: dat

REML criterion at convergence: 71.1

Scaled residuals:
    Min      1Q  Median      3Q     Max 
-1.1417 -0.6147 -0.1494  0.5772  1.3390 

Random effects:
 Groups   Name        Variance Std.Dev.
 blk      (Intercept) 28.35    5.324   
 Residual              5.85    2.419   
Number of obs: 15, groups:  blk, 5

Fixed effects:
            Estimate Std. Error     df t value Pr(>|t|)
(Intercept)   47.200      2.615  5.054  18.047 8.76e-06
trtB           4.600      1.530  8.000   3.007   0.0169
trtC          -1.000      1.530  8.000  -0.654   0.5316
```

What is a Computer Experiment

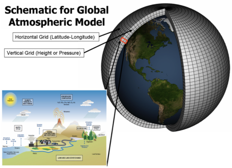
In some situations, it is economically, ethically, or simply not possible to run a **physical experiment**. Instead, the following scenario might be feasible:

- the physical process can be described by a mathematical model (e.g., a system of differential equations)
- computer code (simulator) can be written to compute the response from the mathematical model

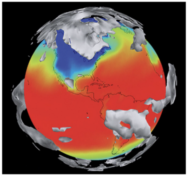
$$\boxed{\begin{array}{c}\text{Input}\\ x \in \mathcal{X}\end{array}} \longrightarrow \boxed{\begin{array}{c}\text{Model}\\ f : \mathcal{X} \mapsto \mathcal{Y}\end{array}} \longrightarrow \boxed{\begin{array}{c}\text{Output}\\ y = f(x)\end{array}}$$

In this case, a researcher can conduct a **computer experiment** by running the computer code, which serves as a proxy for the physical process, to compute a “response” at any combination of values of the inputs

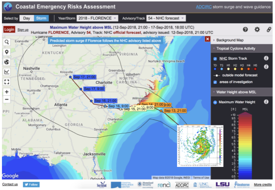
Examples of Computer Models



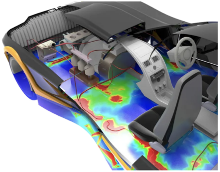
Source: Wikipedia: Climate Model



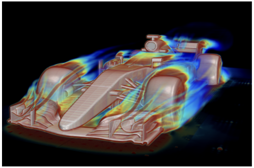
Source: Pacific Northwest National Laboratory



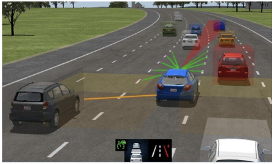
Source: Coastal Emergency Risk Assessment



Source: Ansys, Inc.



Source: Simcenter STAR-CCM+



Source: MATLAB & Simulink

Computer Experiments vs. Physical Experiments

- *"Experimental results are believed by everyone, except for the person who ran the experiment"*

- *"Computational results are believed by no one, except the person who wrote the code"*

> Replication, randomization and blocking are irreverent for a computer experiment because many **computer codes are deterministic** and **all the inputs to the code are known and can be controlled**

Here we are concerning about design and analysis of computer experiments:

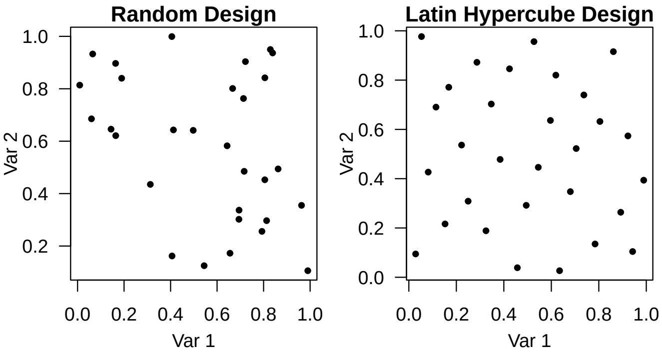
- Design: Which configurations of $\{\boldsymbol{x}_i\}_{i=1}^n$ to run the computer model

- Analysis: How to estimate the input-output relationship $y = f(\boldsymbol{x})$ using data $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ from a computer experiment

Design of Computer Experiments

Question: where to make the runs, i.e., the selection of inputs $\{\boldsymbol{x}_i\}_{i=1}^n$ for a given computational budget n.

Example: $\boldsymbol{x}_i = (x_{i1}, x_{i2})^T \in [0,1]^2$ with $n = 30$



The Latin hypercube design is an example of space-filling designs, which allow for an evenly spread of points encompassing the entire design space.

Analysis of Computer Experiments (aka Emulation)

Goal: fit a statistical model to the computer model inputs-output $\{y_i, \boldsymbol{x}_i\}_{i=1}^n$ to "emulate" the simulator and to quantify the prediction uncertainty for $y(\boldsymbol{x}_{\text{new}})$ via a Gaussian Process Model $\mathrm{GP}\left(m\left(\cdot\right), K\left(\cdot,\cdot\right)\right)$, where

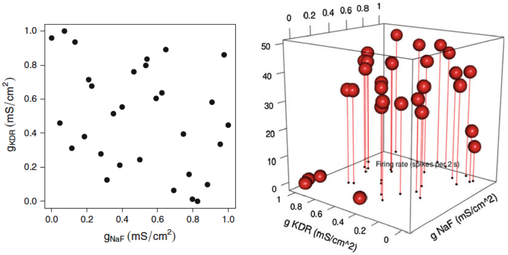
- $m(\boldsymbol{x}) = \mathrm{E}[y(\boldsymbol{x})]$ is the mean function, usually takes a simple form, e.g., $m(\boldsymbol{x}) = \mu$
- $K(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{Cov}(y(\boldsymbol{x}), y(\boldsymbol{x}'))$ is the covariance function, usually parametrized by "distance". e.g., $K(\boldsymbol{x}, \boldsymbol{x}') = C(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\theta}) = \sigma^2 \prod_{j=1}^p C_j(d(x_j, x_j'); \theta_j)$.

Parameters (μ, σ^2, $\{\theta_j\}_{j=1}^p$) can be estimated by fitting a GP model to $\{y_i, \boldsymbol{x}_{i=1}^n\}$ via maximum likelihood method. The prediction (e.g., predicting $y(\boldsymbol{x}_{new})$) and prediction uncertainty can be carried out using the Gaussian conditional distribution formula

Neuron Experiment [pp.776-778, DAE 2017]

The firing rate of a neuron at +380 pA current injection of a young monkey is modeled as a deterministic function of two input variables:

- x_1 gNaF: maximal conductance of the transient sodium
- x_2 gKDR: maximal conductance of the delayed-rectifier potassium



(a) 30 × 2 LHD (b) Firing rates

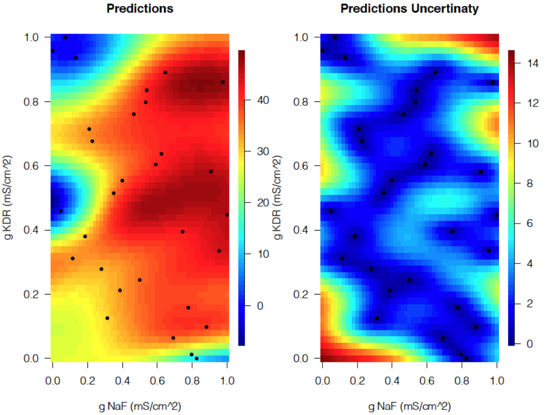
Source: Fig. 20.6, DAE 2017

The goal is to reconstruct the 2D surface within the input space

Neuron Experiment Result

- A GP with squared exponential covariance function (i.e., $C(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 e^{-[\theta_1(x_1 - x_1')^2 + \theta_2(x_2 - x_2')^2}]$) is fitted to $\{y_i, \boldsymbol{x}_i\}_{i=1}^n$ with the estimated parameters $\hat{\mu} = 27.61$ $\hat{\sigma}^2 = 251.86$, $\hat{\theta}_{\text{NaF}} = 5.03$, $\hat{\theta}_{\text{KDR}} = 50.22$.

- With these estimated parameters one can calculate the predictions (**Left**) and their prediction uncertainties (**Right**)



Summary

These slides cover:

- Random and Mixed Effects Models
- Computer Experiments: Concepts, Design and Analysis

`R` functions to know:

- **Random and Mixed Effects Modeling**: `lmer` from the packages `lme4`/`lmerTest`
- **Design and Analysis of Computer Experiments**: `maximinLHS` from the package `lhs` for conducting Latin hypercube sampling designs and `mlegp` from the package of the same name for GP emulation