Lab 11: Randomized Block Design and Nested Design

OBJECTIVES

1. Perform ANOVA for factorial design and randomized block design

2. Nested Design and simple CONTRAST statement

The objective of an experimental design is to provide the maximum amount of reliable information at the minimum cost. In statistical terms, the reliability of information is measured by the standard error of estimates (that is directly related with the population variance, inversely related to sample size). Properly applied experimental design may effectively reduce the population variance, and/or could structure data collection to reduce the magnitude of the experimental error. Usually data resulting from the implementation of experimental designs are described by linear model and analyzed by the analysis of variance as was introduced last week.

Random block design is one of the simplest and probably the most popular experimental design. In this design the sample of experimental units is divided into groups or blocks and then treatments are randomly assigned to units in each block. In some experiments, blocks may be of sufficient size to allow several units to be assigned to each treatment in a block. Such replication of treatments is referred to as randomized blocks with sampling. The linear model for data from such an experiment is

$$y_{ijk} = \mu + \tau_i + \beta_j + \tau_{\beta ij} + \varepsilon_{ijk} \ (i = 1, 2, \dots, t; j = 1, 2, \dots, b; k = 1, 2, \dots, k)$$

Where y_{ijk} is the observed value of the response variable in the kth replicate of treatment i in block j; μ is the overall mean; τ_i is the fixed effect of treatment i; β_j is the effect of block j, a random variable with mean zero and variance σ^2_β; $\tau_{\beta ij}$ is the experimental error, a random variable with mean zero and variance $\sigma^2_{\tau\beta}$; and ε_{ijk} is the sampling error, which is the measure of variation among units treated alike within a block, a random variable with mean zero and variance σ^2.

As you might be aware that in the above model, if the block effect is fixed, the interaction is also fixed so that it becomes a two-factor factorial design and F ratios for all the tests using the sampling error in the denominator.

Nested design refers to some experimental situations where experimental units may contain sampling units, which may, in turn, contain sample subunits. Since the design describes subsamples nested within sample or experimental units, it is called nested or hierarchical design. The linear model for nest design is

$$yijk = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{k(ij)} \ (i = 1, 2, \dots, a; j = 1, 2, \dots, b; k = 1, 2, \dots, n$$

where y_{ijk} is the kth observed value for level i of factor A and level j of factor B which is nested in the ith level of factor A; μ is the overall mean; α_i is the effect of the ith level of factor A; $\beta_{j(i)}$ is the effect of level j of factor B nested in the ith level of factor A; and $\varepsilon_{k(ij)}$ is the variation among sampled units and is the random error.

The subscript j(i) is used to denoted that different j subscripts occur within each value of i; that is, they are "nested" in i. Likewise, the k subscript is "nested" in groups identified by the combined ij subscript (Keep in mind that it is not interaction term).

In this week's lab we will explore the Random Block Design with Sampling and Nested Design with various R tools. Also, simple contrast will be introduced as well.

LABORATORY INSTRUCTIONS

Part I.

Housekeeping Statements

Before we dive into the main part of the code it is good to create a pre-amble in which we will load all the necessary packages for R to execute the following tasks. If you have them installed already great. If not, you can install it using the "packages" tab on the bottom right panel. Click install and put the name of the package you want on the "install packages" window that pops up. The default setting is installing the packages from the CRAN repository where most "mainstream" packages can be found.

To activate the packages, use the following commands:

```
library(easypackages)
libraries("ggplot2","MASS","aod","ResourceSelection","lme4","lmerTest","multcomp","gmodels","car","lsmeans")
```

Nested Design with Subsampling:

The data set of nested design is from Statistical Methods II by Roger Cue of Department of Animal Science from McGill University. This experiment tests the effects of different treatments on the growth of apples. There are three types of treatments which are applied to 12 randomly selected apple trees (four trees per treatment). As you can see the trees are nested in the treatment. At the end of the experiment, six apples were randomly selected from each apple tree and the weight of each apple was recorded. **lmer** will be performed to test the overall treatment effect, and simple CONTRAST will be used to test the difference between the treatments.

First, we need some codes to read the dataset, view it and make sure the categorical variables are treated as such. The following commands will do the trick:

```
nested=read.table("data_lab11.txt", header = TRUE)

nested$tree=as.factor(nested$tree)

nested$apple=as.factor(nested$apple)

nested$treatment=as.factor(nested$treatment)
```

The name of the Data frame is nested. Although the variables tree, apple and treatment have numeric values they are actually factors and we let R know that with the three lines of code following the read table argument.

To fit a nested model we use the command:

```
fit <- lmer(weight ~ treatment +(1|tree/apple),data=nested)
```

lmer is a function of the package lme4 and in the syntax we need to specify that the model is of weight with treatment being the explanatory variable and the fact that the choice of apples is nested under the choice of trees. The last argument of lmer is there to define the dataset, in our case nested.

```
anova(fit)
```

The command above will output the Type III Analysis of Variance Table with Satterthwaite's method for the treatment only of the model above,

```
Type III Analysis of Variance Table with Satterthwaite's method
          Sum Sq Mean Sq NumDF DenDF F value    Pr(>F)
treatment  20747   10374     2    66  82.861 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

whereas the command:

```
rand(fit)
```

will output an ANOVA-like table for random-effects: Single term deletions

```
Model:
weight ~ treatment + (1 | apple:tree) + (1 | tree)
                 npar  logLik    AIC     LRT Df Pr(>Chisq)
<none>              6 -270.10 552.20
(1 | apple:tree)    5 -270.10 550.20 0.00000  1     1.0000
(1 | tree)          5 -270.34 550.67 0.47613  1     0.4902
```

To add a specific contrast of **treatment 1 - (treatment 2+3)/2** we need the code:

```
contest(fit, c(1,-1/2,-1/2), rhs = 0, joint = TRUE,

    collect = TRUE, confint = TRUE, level = 0.95,

    check_estimability = FALSE, ddf = c("Satterthwaite"))
```

The rest of the code conducts some post-hoc analysis to be able to do Tukey Contrasts amongst the three different treatments:

```
posthoc = glht(fit,linfct = mcp(treatment="Tukey"))

mcs = summary(posthoc,test=adjusted("single-step"))

mcs
```

The result is thus:

```
Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lmer(formula = weight ~ treatment + (1 | tree/apple), data = nested)

Linear Hypotheses:
            Estimate Std. Error z value Pr(>|z|)
2 - 1 == 0    12.90        3.23   3.993 0.000208 ***
3 - 1 == 0    40.68        3.23  12.595  < 1e-04 ***
3 - 2 == 0    27.79        3.23   8.602  < 1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

If one wants to see the Tukey grouping the following line will do the trick

cld(mcs,level=0.05,decreasing=TRUE)

In order to get a plot of the 95% family-wise confidence level we need the code:

plot(mcs)

Some plots regarding the residuals can be recovered with the codes:

hist(residuals(fit))

plot(fitted(fit), residuals(fit))

Part II.

Random Block Design with Sampling:

The data set of random block design comes from your textbook (Chapter10, Problem 8, Table 10.30 from Dr. Geaghan's website). An experiment is conducted to test the effectiveness of three types of gasoline additives for boosting gas mileage on a specific type of car. Three randomly selected cars are purchased for the experiment. Each additive is tested four (randomly ordered) times on each of the three cars. In this lab we will perform both a factorial (**Car as fixed effect**) and randomized block design ANOVA (**Car as a random block**) using **lme**. The link of the data is: http://www.stat.lsu.edu/exstweb/statlab/datasets/fwdata97/FW10P08.txt

The variables in the dataset are:
Additive: Type of additive;

Car: ID of cars;
Run: ID of test on each car;
Mpg: Mileage per gallon, the dependent variable.

To load the data we use the line:

```
gas=read.table("data_lab11_set2.txt", header =FALSE)
```

In order to define the variable names, we need to use the colnames argument as follows:

```
colnames(gas)=c("OBS","ADDITIVE","RUN","MPG","CAR")
```

We also need to make sure that the variable additive is recognized as a factor using the following command:

```
gas$ADDITIVE=as.factor(gas$ADDITIVE)
```

A quick side by side boxplot is generated with:

```
p1 <- ggplot(gas, aes(x = ADDITIVE, y = MPG)) +

  stat_boxplot(geom ='errorbar',width = 0.5)+

  geom_boxplot()+

  stat_summary(fun.y=mean, colour="darkred", geom="point",

        shape=18, size=3,show.legend = FALSE) +

  labs(x="ADDITIVE",y="MPG")+

  theme_classic()
p1
```

To create a factorial design, we use:

```
fit2 = lm(MPG ~ ADDITIVE + CAR + ADDITIVE:CAR, data = gas)

Anova(fit2)
```

Followed by a post Anova analysis with

```
lsmeans(fit2,pairwise ~ ADDITIVE, adjust="tukey")     ### Tukey-adjusted comparisons

lsmeans(fit2,pairwise ~ CAR, adjust="tukey")     ### Tukey-adjusted comparisons

lsmeans(fit2,pairwise ~ ADDITIVE:CAR, adjust="tukey")   ### Tukey-adjusted comparisons
```

Then to get the plots with the residuals you use:

```
hist(residuals(fit2))

plot(fitted(fit2), residuals(fit2))
```

In order to create a mixed model now you use:

```
fit3 <- lmer(MPG ~ ADDITIVE + (1 | CAR)+(1|CAR:ADDITIVE), data = gas,REML = FALSE)
```

And the summary will give you the corresponding statistics:

```
summary(fit3)
```

Or the more thorough

```
Anova(fit3)
```

As usual to analyze the residuals you need:

```
hist(residuals(fit3))
plot(fitted(fit3), residuals(fit3))
```

LAB ASSIGNMENT
Your assignment is to perform necessary analysis using SAS and answer the following questions (Please do not print all the output. Only print the graphs and tables that you think are relevant to your answers).
Part I. Answer the following questions for data set of nested design:

1. According to the model, do different treatments have significant effect on the apple weight?
2. Using the contrast to test the hypothesis that the difference between treatment 1 and the average of treatment 2 and treatment 3 is equal to zero. What is your conclusion?

Part II. Answer the following questions for data set of random block design:

1. Take the variable Car as a random block. Do different additives have significantly different effect on boosting gas mileage?
2. Take the variable Car as a fixed effect. Do different additives have significantly different effect on boosting gas mileage?
3. Are your answers for question 1 and 2 consistent? If not, why (that is, why the test statistic and p-value are different)?
4. Which model is more appropriate? Why?
5. What is the additional assumption we must have for the RBD model? Can we test it?

*Remember to attach your R code log with your lab report