Latin Square Design and Expected Mean Square

Lab 12 R Notes: EXST 7014/15

Contents

0.1 Objectives

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.

In last week's lab, a special RBD example was exercised, which help you understand why it is important to consider the block effect in the experimental design. You are also getting familiar with how to construct source ANOVA table with variance sources, their degree of freedom and expected mean squares. Today, Latin Square Design (LSD) with row and column blocks will be introduced. Compared with factorial design, LSD is more efficient which allows the researchers to control variations in two directions with minimum experimental unit. In LSD, there are equal numbers of treatment, row block and column block. Treatments are assigned at random within rows and columns, with each treatment only once per row and only once per column. Two blocks of column and row are in a factorial structure. The linear model for data from such an experiment is

$$Y_{ij} = \mu + \alpha_i + \beta_j + \tau_k + \epsilon_{ijk} \quad (i = 1, 2, ..., r: \quad j = 1, 2, ..., r: \quad k = 1, 2, ..., r)$$

Where $\alpha_{\mathbf{i}}$ is the effect of row block; $\beta_{\mathbf{j}}$ is the effect of column block, $\mathbf{y_{ij}}$ is the observed value of the responsible variable in the $\mathbf{i^{th}}$ row of the column $\mathbf{j}$. μ is the overall mean; $\tau_{\mathbf{k}}$ is the fixed effect of treatment $\mathbf{k}$; $\epsilon_{\mathbf{ijk}}$ is the random error with the mean zero and variance $\sigma^{\mathbf{2}}$. Note that $\mathbf{i} = \mathbf{j} = \mathbf{k}$ for any Latin Square.

The null hypothesis test of one-way ANOVA is that the means of the response variable are the same for the different levels of treatment ($H_0 : \mu_1 = \mu_2 = \ldots = \mu_t$); the alternative hypothesis is that they are not all the same.

There are three assumptions need to be considered for ANOVA: the treatments are independently sampled; residuals or deviation of observations within groups should be normally distributed (evaluated by residual plot, normality test); and the variance from each level of treatment is the same (i.e. homogeneous variance, Bartlett test is preferred method to evaluate the homogeneity of variance).

The Latin Square Design is appropriate only if effects of all three factors (row block, column block and treatment) are additive, i.e., all interactions are zero. It is a very important assumption of Latin Square Design. The ANOVA table of LSD is as the following:

Source	DF	Expected Mean Square (EMS)
Treatment	$r-1$	$\sigma^2 + r\sigma_\tau^2$
Row	$r-1$	$\sigma^2 + r\sigma_\alpha^2$
Column	$r-1$	$\sigma^2 + r\sigma_\beta^2$
Error	$(r-1)(r-2)$	σ^2
Total	r^2-1	

0.2 Lab Setup

Run the following code to both install and load the required packages.

```
#' a function to only install needed but unavailable packages
#' and loads these packages after installation

ipak <- function(pkg){
    new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
    if (length(new.pkg))
        install.packages(new.pkg, dependencies = TRUE)
    sapply(pkg, require, character.only = TRUE)
}

# use function to install and load packages
packages <- c('lme4', 'EMSaov')
ipak(packages)


#' lme4       # to fit the linear mixed-effect model
#' EMSaov     # to extract expected mean square(EMS) formulae for the ANOVA components
```

0.3 The Data

The data set is from Statistics for Experimenters by Box, Hunter and Hunter (Chapter 8). An experiment was carried out to compare four gasoline additives (A, B, C and D) for reducing NOX emissions from cars. Four drivers (I, II, III and IV) and four cars (1, 2, 3 and 4) will be used in the experiment. Because there may be significant driver effects and car effects, we would like to block on both variables. A Latin square design accomplishes this. In this design, each additive is used exactly once by each driver, and exactly once in each car. However, not every combination of driver and car is used with each additive. The design can be visualized as a square with four rows corresponding to the drivers and four columns corresponding to the cars. Each additive occurs exactly once in each row and exactly once in each column.

```
#' Create an object called @auto to host the data set
#'
#' @colClasses= to specify the respective data classes of each column
#'    N/B the 1st 3 columns are factor sor repeat (@rep) "factor" 3 times
#'    the last column is an integer so, "integer"

auto <- read.table(header = T,
                   colClasses = c(rep("factor", 3), "integer"),
```

```r
                    text=' car driver additive reduct
                           1  I    A  21
                           2  I    B  26
                           3  I    D  20
                           4  I    C  25
                           1  II   D  23
                           2  II   C  26
                           3  II   A  20
                           4  II   B  27
                           1  III  B  15
                           2  III  D  13
                           3  III  C  16
                           4  III  A  16
                           1  IV   C  17
                           2  IV   A  15
                           3  IV   B  20
                           4  IV   D  20 ')
str(auto)
```

```
## 'data.frame':    16 obs. of  4 variables:
##  $ car     : Factor w/ 4 levels "1","2","3","4": 1 2 3 4 1 2 3 4 1 2 ...
##  $ driver  : Factor w/ 4 levels "I","II","III",..: 1 1 1 1 2 2 2 2 3 3 ...
##  $ additive: Factor w/ 4 levels "A","B","C","D": 1 2 4 3 4 3 1 2 2 4 ...
##  $ reduct  : int  21 26 20 25 23 26 20 27 15 13 ...
```

0.4 Fitting the Model

car and driver are assumed to be random by the nature of the experiment, hence we need to run a mixed-effects model. As such we are going to use the **lmer function** (for Linear Mixed and multilevel modelling - from the lme4 package).

```r
#' Get exhaustive notes with examples on fitting mixed effects models in R
#'    Search Ouptput would be in the Help Subpane in RStudio
vignette(topic="lmer", package = "lme4")
```

0.4.1 Standard Model fitted with Restricted (Residual) Maximum Likelihood Estimation

The **lmer** function, by default, estimates the variance-covariance components using the REML estimation technique, and employs the satterwaithe method in computing the respective denominator degrees of freedom of each component.

```r
#' The model object is named to mirror the different models fitted using SAS
#' random effects are specified with the pipe symbol sandwiched between
#'     1 and the name of the random effect variable e.g. (1|car)


fit.REML <- lmer(reduct ~ additive + (1|car) + (1|driver), data=auto)
summary(fit.REML)
anova(fit.REML)   # Get Type 3 Anova Test
```

0.4.2 Model with EMS Details

To get the expected means square computation formula for each component in the ANOVA source table we would resort to the **EMSanova** function from the **EMSaov** package

```r
#' "additive" is treated as a fixed effect,
#'    "car" and "driver" are treated as random effects.
#'     Therefore, @type = c("F","R","R") following their order in the formula
fit.Type3 <- EMSanova(reduct ~ additive+car+driver, data=auto,
                      type=c("F","R","R"))

#' Remove or Pool interaction (:) effects to the error/residual term
del.ID <- c("additive:car", "additive:driver", "car:driver", "Residuals")
fitType3_Pooled <- PooledANOVA(fit.Type3, del.ID)

fitType3_Pooled
```

	Df	SS	MS	Fvalue	Pvalue	Sig	EMS
additive	3	40	13.3333	5	0.0452	*	Error+16additive
car	3	24	8.0000	3	0.117		Error+16car
driver	3	216	72.0000	27	7e-04	***	Error+16driver
Residuals	6	16	2.6667				Error

EMSanova automatically fits two-way interactions between the model terms - **additive, driver and car**. Since our model does not assume any such relationships (interaction), we need to remove all such effects or reallocate them as unexplained error. The **PooledANOVA** function is used to pool (add) or reallocate all these unneeded interaction effects to the error term. It (**PooledANOVA**) takes 2 arguments:

1. an object that hosts the original EMSanova function (in our case, **fit.Type3**)

2. a character vector with the ID (or R names) of the various interactions (see, **del.ID** above)

```r
#' Get quick help with examples on the EMSanova and Pooled ANOVA functions
?EMSanova
?PooledANOVA
```

0.4.3 CAUTION on Reading the EMS in R

Let a be a constant and X be a random variable, then $Var(aX) = a^2Var(X)$.

For some reason **SAS** reports $\mathbf{aVar(X)}$, whereas **R** reports $\mathbf{a^2Var(X)}$ in their respective EMS columns.

0.5 Checking Assumptions

0.5.1 Test for Homogeneity of Variance

```r
# Test for Homogeneity of Variance
bartlett.test(reduct ~ additive, data=auto)
```

0.5.2 Testing for Normality

```
#' (Residuals vs. Fitted Values) and QQ Plots colored by Additive types
plot(fit.REML, col=auto$additive, which=c(1,2))

#' Q-Q Plot with Confidence Bands
#require(MASS)
qqp(residuals(fit.REML), "norm")

#' Alternative QQ-Plot
qqnorm(residuals(fit.REML))
qqline(residuals(fit.REML))
```

```
#' Test for Normality of Residuals
shapiro.test(residuals(fit.REML))
```

0.6 Lab Assignment

Your assignment is to perform necessary analysis using either SAS or R to answer the following questions Only print the graphs and tables that you think are relevant to your answers.

0.6.1 Question 1

According to the model, does the additive type have any significant effect on reducing NOX emission? Can the effects of Cars or Drivers be ignored in this model?

0.6.2 Question 2

There is an important assumption for Latin Square Design besides the usual assumptions for the linear model (normality, homogeneous variance et al). What is that?