# Notes and Solved examples from 'Design and Analysis of Experiments with R by John Lawson'

May 14, 2022

## 1 INTRODUCTION

The book 'Design and Analysis of Experiments with R' by John Lawson is an excellent open source tool which has helped me to accelarate my knowledge in statistics and use of R to propose & analyze experimental results. In this paper I have shared key learnings from the book. To demonstrate my understanding, I have solved a couple of examples from each chapter.

## 2 C1: Completely randomised design with one factor

The first chapter is a gentle introduction to design with single factors. A single factor can be, for example, yeast (bread making), baking powder (biscuit making), catalyst (chemical reaction) etc. Level defines the amount of factor. Thus, in the case of bread making, the amount of yeast for a given amount of dough can vary and this variation is referred to as 'level'. Factor is an independent variable and the level affects a certain property of the experimental subject. Thus for example, in the case of bread making, the property of interest could be the height of the bread. Our objective is to know how the level of yeast impacts the height. A statistician would design experiments with different levels of yeast and collect data for the height of the bread reached in a controlled environment to understand the relationship between yeast level and height. It is apparent that height is a dependent variable. The key to a good design is (a) to ensure that there are sufficient amount of genuine repeats for each level. This is necessary to compute the variability in the dependent variable (the height for bread example) (b) Randomisation of experiments to rule out bias in experimentation caused by external factors.

CRD mathematical model with unequal number of each factor is

$$Y = \mu_i + \epsilon \tag{1}$$

Where $mu_i$ is the mean for the specific level of the treatment factor (e.g. mean height of the heights of loafs treated at $40^oC$). $epsilon$ is the distribution of the experimental errors

or...

$$Y = \mu + \tau + \epsilon \tag{2}$$

Where $\tau$ are called the effects. They represent the difference between the long-run average of all possible experiments at the level of treatment factor and the overall average (in other words, the difference between the average height for all the loafs allowed to rise for specific temperature group and the average overall height). The $mu$ is for overall average (overall mean). The mathematical formulation (matrix representation) for a linear model based on single factor is as follows

$$SSE = y'y - \hat{\beta}X'y..SST = y'y - (1'y)^2/(1'1)..SSR = \hat{\beta}'X'y - (1'y)^2/(1'1)..SST = SSR + SSE..e = y - \hat{y} \tag{3}$$

$$\hat{\beta} = (X'X)^{-1}X'y \tag{4}$$

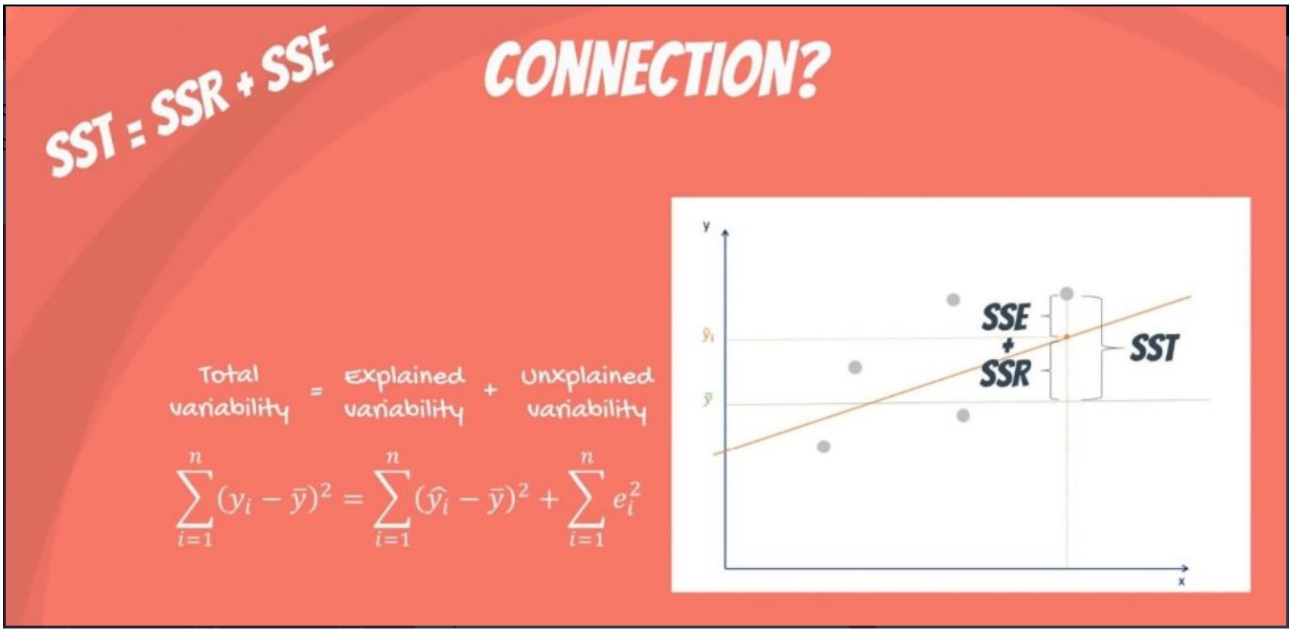
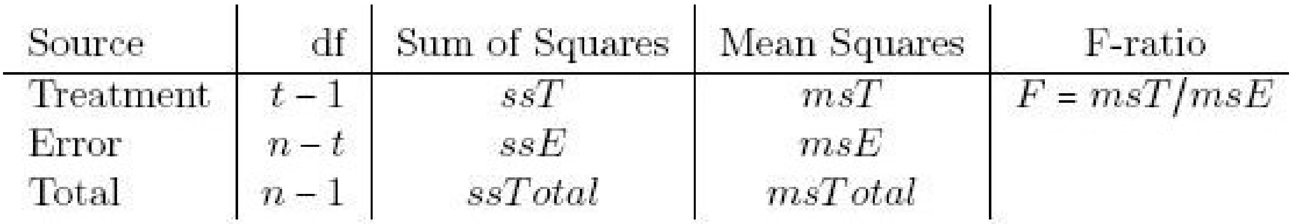

SST = SSR + SSE
CONNECTION?
Total variability = Explained variability + Unxplained variability
$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2$
y
x
$\hat{y}_i$
$\bar{y}$
SSE
+
SSR
SST

Figure 1: Graphical presentation of SSR and SSE (components of total variability, SST). Source: 365datascience.com

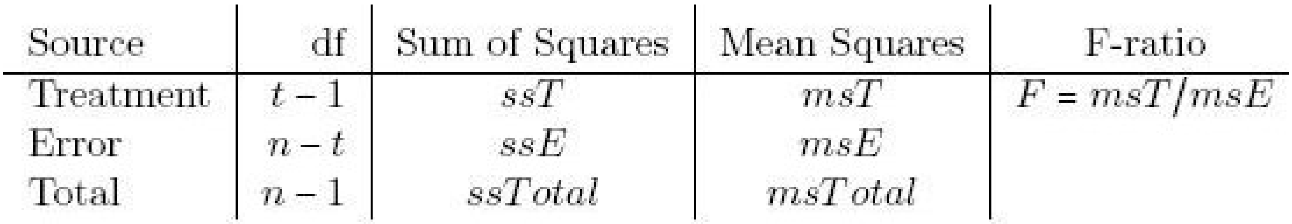


| Source | df | Sum of Squares | Mean Squares | F-ratio |
|---|---|---|---|---|
| Treatment | $t-1$ | $ssT$ | $msT$ | $F = msT/msE$ |
| Error | $n-t$ | $ssE$ | $msE$ | |
| Total | $n-1$ | $ssTotal$ | $msTotal$ | |

Figure 2: Graphical presentation of SSR and SSE (components of total variability, SST). Note Source: 365datascience.com

## Hypothesis test of no treatment effects

In the CRD model, the hypotheses of interest are $H_0$ (where the null hypothesis is supported if $\mu_1 = \mu_2 =$ ... $\mu_t$ or same for the effects($\tau$)) versus $H_a$ (where atleast two of the $\tau$ differ). If $H_0$ is supported (i.e. means of the various treatment factors equal to each other), then both SSR and SSE follow the chi-squared distribution. These sums of squares are presented in the ANOVA table with their corresponding Mean Squares (achieved by dividing by their appropriate degrees of freedom) as shown below.



| Source | df | Sum of Squares | Mean Squares | F-ratio |
|---|---|---|---|---|
| Treatment | $t-1$ | $ssT$ | $msT$ | $F = msT/msE$ |
| Error | $n-t$ | $ssE$ | $msE$ | |
| Total | $n-1$ | $ssTotal$ | $msTotal$ | |

Figure 3: Analysis of Variance of Table. Note, $ssTotal$ refers to SST in the text and $ssT$ refers to SSR

Under $H_0$, the F-ratio follows the F-distribution with t-1 and n-t degrees of freedom. $H_a$ follows the noncentral F distribution with noncentrality parameter

One-way analysis of variance (ANOVA) is used to understand if the effect of factor levels is statistically significant. The test statistics is essentially the ratio of mean squared error of regression (MSR) to mean squared error of residuals (MSE) and follows F-distribution. One of the conditions that should be satisfied is that errors are homogenous or equal across different levels of treatment. When this is not the case, following four ways are suggested to regularise the variance

- Transforming the response variable using Box-Cox transformation
- Weighted linear model
- Transforming the response based on distribution of the response

A second condition is errors are from standard normal distribution (mean=0, sd=1). Checking this condition is met can be done qualitatively through QQ plot and histogram of errors. Quantitative method, not discussed in the first chapter, is Lilliefors test for normality. The starting point for ANOVA test is that null hypothesis is true. When this is not the case the test statistics does not follow F-distribution but rather a wider spread. The probability of rejecting a null hypothesis when it is not true is called Power. Higher the power, higher the probability of correctly rejecting null hypothesis when it is not true. It is desirable to reach a high power which is directly proportional to the sample size or number of experiments. Lastly, the chapter discusses on post-hoc comparison of treatment. Pre-planned and un-planned comparisons are discussed.

**I have solved exercise 4 on page 52 of this chapter based on the concepts discussed above. The solution is stored my Github**.