STAT 210
Applied Statistics and Data Analysis
Week 7 - Summary

Joaquin Ortega

King Abdullah University of Science and Technology

Announcements

- First exam will be on Saturday, October 22, 9:00 - 12:00 am, Room 2322.
- The exam is based on R. You will need to bring your computers.
- You can use the notes, presentations, books and exercises we have solved in the class, but you are not allowed to use resources from the internet outside KAUST.
- The exam will be posted in Blackboard and you have to submit your solution through Blackboard by 12:00.
- You need to submit two documents, a pdf with your answers, and a script with the R code. The script can be a Rmarkdown file.

Videos 22 and 23: Analysis of Variance

Analysis of Variance

Analysis of Variance (Anova) is a statistical tool to compare several means, which are related to the levels of the different factors.

This allows us to differentiate the effects of several factors that are varied together in a single experiment.

However, to do this, we need to analyze the variances associated with the different means.

One-way Analysis of Variance

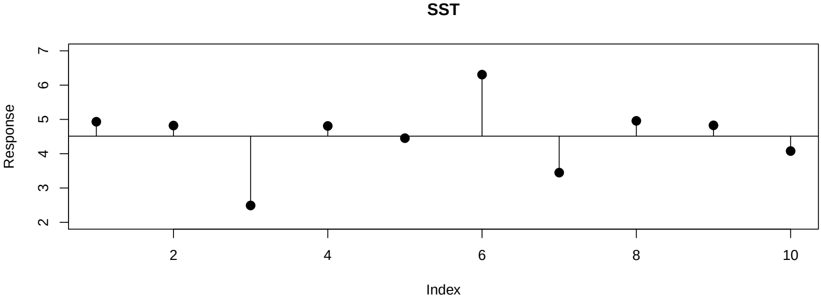


Figure 1: Outcomes of ten replications of the experiment.

The horizontal line corresponds to the overall mean $\bar{y}_{\bullet\bullet}$ for the response. The vertical segments are the differences between the observed values and the average,

$$y_{ij} - \bar{y}_{\bullet\bullet},$$

for $i = 1, 2, j = 1, \ldots, 5$.

One-way Analysis of Variance

If we color the points in the graph according to the level of the factor, we get

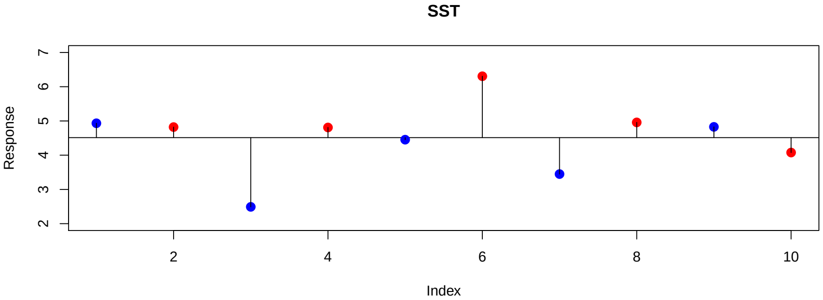


Figure 2: Total sum of squares. Colors correspond to the two treatment levels.

Next, let us consider the data for each level and calculate a separate mean, $\bar{y}_{i\bullet}$ for the values corresponding to $L_i, i = 1, 2$. The graph is

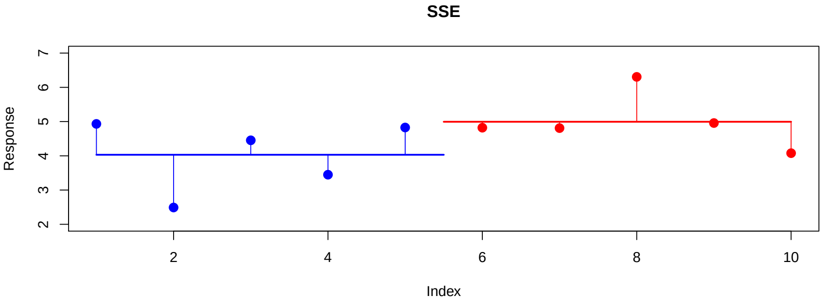


Figure 3: Error sum of squares.

One-way Analysis of Variance

The **total sum of squares** SST is defined as

$$SST = \sum_{i=1}^{2} \sum_{j=1}^{5} (y_{ij} - \bar{y}_{\bullet\bullet})^2.$$

The sum of the squares of the differences between the observed values and the corresponding treatment mean is known as the **error sum of squares**

$$SSE = \sum_{j=1}^{n_1} (y_{1j} - \bar{y}_{1\bullet})^2 + \sum_{j=1}^{n_2} (y_{2j} - \bar{y}_{2\bullet})^2.$$

SSE is the sum of the squares with respect to the red and blue lines in figure 3.

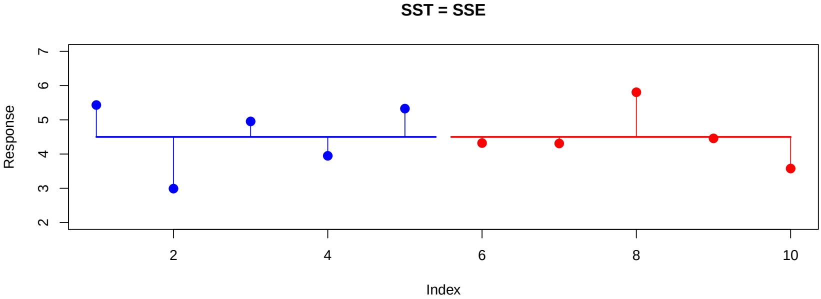


Figure 4: Total and error sum of squares.

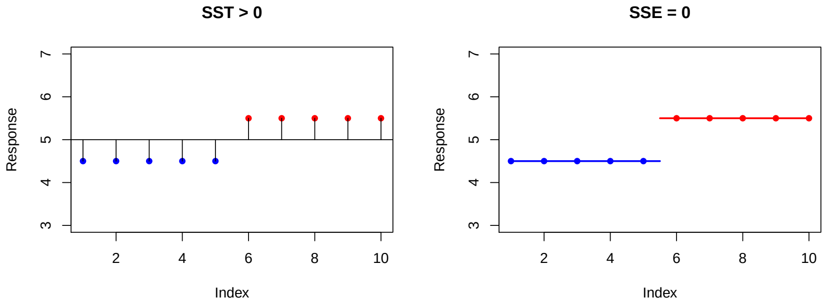


Figure 5: Total and error sum of squares.

One-way Analysis of Variance

$$SST = SSE + \sum_{i=1}^{2} n_i(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2. \tag{1}$$

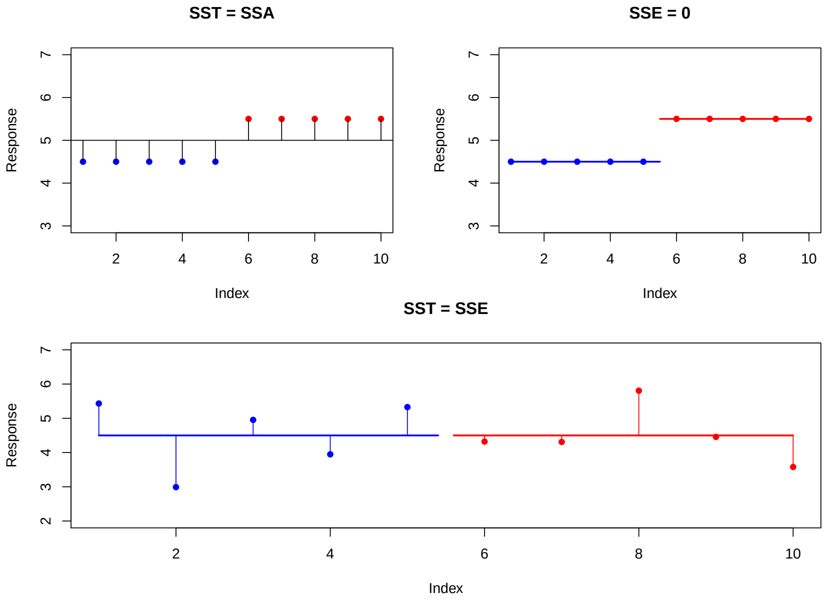
Since the second sum in (1) is positive, we see that $SSE \leq SST$.

The difference between SST and SSE is known as the **treatment sum of squares** and will be denoted by SSA:

$$SSA = \sum_{i=1}^{2} n_i(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2.$$

When differences in mean between treatments are significant, SSA will be big with respect to SSE. If, on the contrary, the means are similar, then SSA will be small with respect to SSE.

One-way Analysis of Variance



One-way Analysis of Variance

SSE represents the variability *within* each group or level of the treatment factor, while SSA represents the variability *between* different groups or factor levels.

To estimate variances, we divide sums of squares by the corresponding degrees of freedom (d.f.).

In general, if we had k treatment levels, we would have $k-1$ degrees of freedom for treatments.

If each factor level were replicated r times, then there would be $r-1$ degrees of freedom for each level, since we lose one for each treatment mean. Considering that there are k levels, there are $k(r-1)$ d.f. for error in the whole experiment.

Finally, the total number of data points in the experiment is $n = rk$, and we lose one for the overall mean, so there are in total $rk-1$ d.f.

One-way Analysis of Variance

The usual way to sum up these results is through an Analysis of Variance (Anova) table.

Table 1: Anova table for example 1.

Source	SS	d.f.	MS	F_{obs}	p-value
Treatment	SSA	$k-1$	$MSA = \frac{SSA}{k-1}$	$F_{obs} = \frac{MSA}{MSE}$	$P(F_{k-1,k(r-1)} > F_{obs})$
Error	SSE	$k(r-1)$	$MSE = \frac{SSE}{k(r-1)}$		
Total	SST	$kr-1$			

A Worked-Out Example

```r
results <- read.table('yields.txt',header=T)
attach(results); str(results)
```

```
## 'data.frame':    10 obs. of  3 variables:
##  $ sand: int  6 10 8 6 14 17 9 11 7 11
##  $ clay: int  17 15 3 11 14 12 12 8 10 13
##  $ loam: int  13 16 9 12 15 16 17 13 18 14
```

```r
frame <- stack(results)
names(frame) <- c('yield','soil');str(frame)
```

```
## 'data.frame':    30 obs. of  2 variables:
##  $ yield: int  6 10 8 6 14 17 9 11 7 11 ...
##  $ soil : Factor w/ 3 levels "sand","clay",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```r
attach(frame)
```

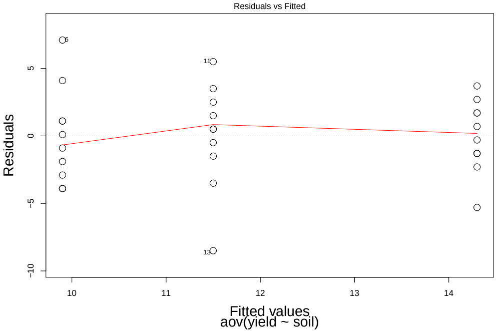
A Worked-Out Example

```r
summary(aov(yield~soil))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## soil         2   99.2   49.60   4.245  0.025 *
## Residuals   27  315.5   11.69
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

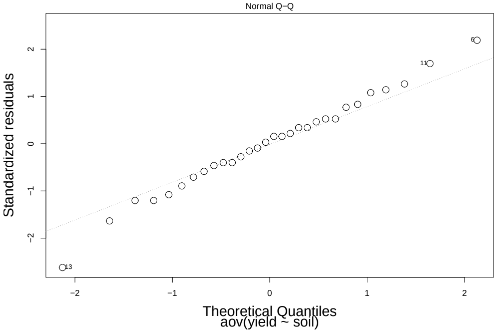
A Worked-Out Example

```
plot(aov(yield~soil), which = 1,  cex=1.7,
     cex.lab=1.7, cex.sub=1.7, cex.main=1.7)
```



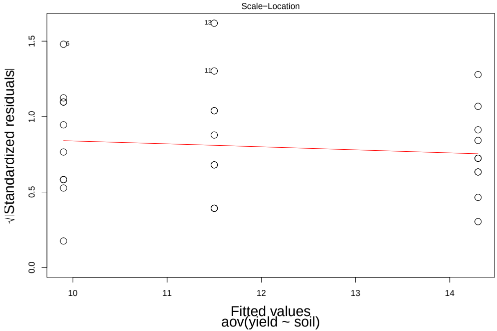
A Worked-Out Example

```
plot(aov(yield~soil), which = 2,  cex=1.7,
     cex.lab=1.7, cex.sub=1.7, cex.main=1.7)
```



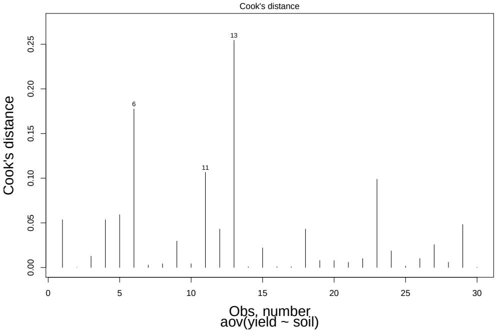
A Worked-Out Example

```
plot(aov(yield~soil), which = 3,  cex=1.7,
     cex.lab=1.7, cex.sub=1.7, cex.main=1.7)
```



A Worked-Out Example

```
plot(aov(yield~soil), which = 4,  cex=1.7,
     cex.lab=1.7, cex.sub=1.7, cex.main=1.7)
```



Video 24: Experimental Design

A General Model for One-way Anova.

A General Model for One-way Anova.

Suppose we have an experiment where one factor or treatment T has k levels and there are r_i replications for configuration i, $i = 1, \ldots, k$.

By this we mean that for each configuration, the experiment is repeated r_i times, and the results obtained are the replications.

It does not mean that the experiment is performed once, and r_i measurements are taken.

A General Model for One-way Anova.

Let y denote the response variable, then a model for this experiment would be

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}; \qquad j = 1, \ldots, r_i, i = 1, \ldots k, \tag{2}$$

where

- y_{ij} represents the outcome of the j-th replication of level i,
- μ is the overall mean,
- τ_i is the **effect** of treatment i and
- ϵ_{ij} represents the error for the j-th replication of level i.

A General Model for One-way Anova.

The errors are assumed to be independent and normally distributed with mean 0 and equal variance σ^2.

In this model, μ is the global mean for the experiment, and $\mu + \tau_i$ represents the average response for the i-th group.

(2) is usually known as the *effects model*.

The total number of data points is $n = \sum_{i=1}^{k} r_i$.

A General Model for One-way Anova.

An alternative model for this experiment would be

$$y_{ij} = \mu_i + \epsilon_{ij}; \qquad j = 1, \ldots, r_i, i = 1, \ldots k, \tag{3}$$

where μ_i represents the average response for level i of the treatment factor and the same assumptions are made for the errors ϵ_{ij}.

Comparing μ_s with μ_t is equivalent to comparing τ_s with τ_t:

$$\mu_t - \mu_s = (\mu + \tau_t) - (\mu + \tau_s) = \tau_t - \tau_s$$

Model (3) is sometimes known as the *cell means model*. The two models are equivalent.

A General Model for One-way Anova.

These models can be used in two different scenarios.

When the experimenter specifically chooses the treatments, and there is no desire to extend the results to other treatments, the model is referred to as a **fixed effects model**.

When the treatments are selected at random from a larger population of possible treatments and the experimenter would like to extend the conclusions of the experiment to all treatments in the population, the model is called a **random effects model**.

We will only consider the fixed effects model.

Least Squares Estimation

Least Squares Estimation

The least squares estimators for the one-way Anova model are the values for the parameters $\hat{\mu}, \hat{\tau}_1, \ldots, \hat{\tau}_k$ that minimize the error sum of squares

$$\sum_{i=1}^{k} \sum_{j=1}^{r_i} \epsilon_{ij}^2 = \sum_{i=1}^{k} \sum_{j=1}^{r_i} (y_{ij} - \mu - \tau_i)^2. \tag{4}$$

The resulting model $y_{ij} = \hat{\mu} + \hat{\tau}_i$ is the best-fitting model in the sense of minimizing (4).

Least Squares Estimation

The procedure for minimizing this expression is the usual. The expression in (4) is differentiated with respect to the parameters $\mu, \tau_1, \ldots, \tau_k$ in turn and each of the resulting expressions is set equal to zero, yielding a set of $k + 1$ equations. These are known as the *normal equations*.

It is an exercise in calculus to verify that these equations are

$$y_{\bullet\bullet} - n\hat{\mu} - \sum_{i=1}^{k} r_i \hat{\tau}_i = 0, \tag{5}$$

$$y_{i\bullet} - r_i(\hat{\mu} + \hat{\tau}_i) = 0, \quad i = 1, \ldots, k, \tag{6}$$

where the hat notation indicates that these are the values that minimize (4).

Least Squares Estimation

From (6) we get that

$$\hat{\mu} + \hat{\tau}_i = \frac{1}{r_i} y_{i\bullet} = \bar{y}_{i\bullet}$$

for $i = 1, \ldots, k$, so the least squares estimate for the i-th treatment mean is the corresponding sample mean $\bar{y}_{i\bullet}$.

However, there is a problem with the normal equations. If we add up the equations in (6) we get (5). The k equations in (6) are linearly independent, but if we add (5) we get an undetermined system of equations that does not have a unique solution.

This means that the $k + 1$ parameters in model (2) are not all estimable.

Least Squares in Matrix Notation

Least Squares in Matrix Notation

Consider model (2) with $k = 3$ factor levels and $r = 2$ replicates for each level. We can write the effects model using matrix notation as

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon} \tag{7}$$

where

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \quad \beta = \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{pmatrix}.$$

Least Squares in Matrix Notation

The least squares estimators are the solution to the normal equations $\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{y}$. The problem is that the matrix $\mathbf{X}'\mathbf{X}$ is singular and cannot be inverted.

The `R` function `lm` makes the matrix $\mathbf{X}$ full rank by dropping the column that corresponds to the first level of the factor:

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}.$$

Least Squares in Matrix Notation

This coding makes the first level of the treatment the standard, and all other levels are compared to it.

For example, with $k = 3$ levels the solution to the normal equations is

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \hat{\beta} = \begin{pmatrix} \hat{\mu} + \hat{\tau}_1 \\ \hat{\tau}_2 - \hat{\tau}_1 \\ \hat{\tau}_3 - \hat{\tau}_1 \end{pmatrix}. \tag{8}$$

Least Squares in Matrix Notation

Parameter	Estimator
μ	$\overline{Y}_{\bullet\bullet}$
μ_i	$\overline{Y}_{i\bullet}$
τ_i	$\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}$
ϵ_{ij}	$Y_{ij} - \overline{Y}_{i\bullet}$
σ^2	$\frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{r_i}(Y_{ij} - \overline{Y}_{i\bullet})^2$

Example

Example

A tire manufacturer is interested in investigating the braking performance for different types of tread patterns.

There are four different tread patterns identified with the letters `A`, `B`, `C`, and `D`. Six measurements were taken with each one.

Measurements (`StopDist`) correspond to the braking distance in feet of a medium sized car from a speed of 60 miles per hour.

The same driver and car were used for all the experiments.

The order of the treatments was assigned at random.

Example

```r
mod0 <- lm(StopDist ~ tire, data = Tire)
anova(mod0)
```

```
## Analysis of Variance Table
##
## Response: StopDist
##           Df Sum Sq Mean Sq F value   Pr(>F)
## tire       3 5673.1 1891.04  5.3278 0.007316 **
## Residuals 20 7098.8  354.94
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example

```r
summary(mod0)
```

```
##
## Call:
## lm(formula = StopDist ~ tire, data = Tire)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.333  -9.667  -2.250  11.417  36.667
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  379.667      7.691  49.363  < 2e-16 ***
## tireB         25.500     10.877   2.344 0.029497 *
## tireC         42.000     10.877   3.861 0.000973 ***
## tireD         30.667     10.877   2.819 0.010594 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.84 on 20 degrees of freedom
## Multiple R-squared:  0.4442, Adjusted R-squared:  0.3608
## F-statistic: 5.328 on 3 and 20 DF,  p-value: 0.007316
```

Example

To interpret these results recall from (8) that the default coding for treatment in `R` means that the first value (`Intercept`) corresponds to the average value for the first treatment level, $\hat{\mu} + \hat{\tau}_1$ while `tireB`$= \hat{\tau}_2 - \hat{\tau}_1$, `tireC`$= \hat{\tau}_3 - \hat{\tau}_1$, and `tireD`$= \hat{\tau}_4 - \hat{\tau}_1$. Thus

$$\begin{aligned}
\hat{\mu} + \hat{\tau}_1 &= 379.7; \\
\hat{\mu} + \hat{\tau}_2 &= 379.7 + 25.5 = 405.2 \\
\hat{\mu} + \hat{\tau}_3 &= 379.7 + 42.0 = 421.7 \\
\hat{\mu} + \hat{\tau}_4 &= 379.7 + 30.7 = 410.4
\end{aligned}$$

.

Variance Estimation

Variance Estimation

The **residuals** $\hat{\epsilon}_{ij}, j = 1 \ldots, r_i, i = 1, \ldots, k$ are defined as

$$\hat{\epsilon}_{ij} = y_{ij} - \bar{y}_{i\bullet} = y_{ij} - (\hat{\mu} + \hat{\tau}_i)$$

and represent the difference between the j-th replication of the i-th treatment and the estimated treatment mean $\hat{\mu} + \hat{\tau}_i = \bar{y}_{i\bullet}$.

The sum of squares for error or error sum of squares is

$$\begin{aligned} SSE &= \sum_i \sum_j \hat{\epsilon}_{ij}^2 = \sum_i \sum_j (y_{ij} - \bar{y}_{i\bullet})^2 \qquad (9) \\ &= \sum_i \sum_j y_{ij}^2 - \sum_i r_i \bar{y}_{i\bullet}^2. \end{aligned}$$

An unbiased estimator for the variance is given by

$$\hat{\sigma}^2 = \frac{SSE}{n-k} = MSE \qquad (10)$$