

MATH 204: PRINCIPLES OF STATISTICS 2
WINTER 2026
THE IMPORTANCE OF BALANCE IN ANOVA

The potato damage data is an **unbalanced** three-way factorial design: the three factor predictors have two levels

- **potato**: variety of potato
- **regime**: preparatory regime
- **temp**: temperature at which the experiment is carried out.

The outcome variable is ion leakage (a measure of damage).

```
> Potato<-as.data.frame(read.spss('../R/DataSets/Potato.sav'))
> Potato.counts<-xtabs(~potato+regime+temp, data=Potato)
> ftable(Potato.counts)

+               temp  -4 C  -8 C
+ potato      regime
+ Variety 1 Room Temp      5    5
+           Cold Room     12   13
+ Variety 2 Room Temp     13   13
+           Cold Room      7    7

> levels(Potato$potato)<-c('V1', 'V2')
> levels(Potato$regime)<-c('Room', 'Cold')
> levels(Potato$temp)<-c('-4', '-8')
```

If we consider the main effects only model, this can be fitted in R as follows

```
> summary(lm(leak~potato+regime+temp, data=Potato))

+
+ Call:
+ lm(formula = leak ~ potato + regime + temp, data = Potato)
+
+ Residuals:
+    Min       1Q   Median       3Q      Max
+ -17.60  -6.38  -1.24   4.60  23.14
+
+ Coefficients:
+              Estimate Std. Error t value Pr(>|t|)
+ (Intercept)      9.95       2.73    3.64  0.00051 ***
+ potatoV2         9.63       2.50    3.85  0.00025 ***
+ regimeCold     -11.03       2.50   -4.42  3.5e-05 ***
+ temp-8          8.19       2.32    3.52  0.00075 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 10.1 on 71 degrees of freedom
+ Multiple R-squared:  0.48, Adjusted R-squared:  0.458
+ F-statistic: 21.8 on 3 and 71 DF, p-value: 3.99e-10
```

The fit is identical if the terms are rearranged, as the model is the same.

```
> summary(lm(leak~regime+potato+temp,data=Potato))

+
+ Call:
+ lm(formula = leak ~ regime + potato + temp, data = Potato)
+
+ Residuals:
+     Min       1Q   Median       3Q      Max
+ -17.60  -6.38  -1.24   4.60  23.14
+
+ Coefficients:
+             Estimate Std. Error t value Pr(>|t|)
+ (Intercept)      9.95       2.73   3.64 0.00051 ***
+ regimeCold     -11.03       2.50  -4.42 3.5e-05 ***
+ potatoV2        9.63       2.50   3.85 0.00025 ***
+ temp-8          8.19       2.32   3.52 0.00075 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 10.1 on 71 degrees of freedom
+ Multiple R-squared:  0.48, Adjusted R-squared:  0.458
+ F-statistic: 21.8 on 3 and 71 DF, p-value: 3.99e-10
```

However, the results of a standard ANOVA are **different** if the terms are rearranged.

```
> print(anova(lm(leak~potato+regime+temp,data=Potato)),digits=6)

+ Analysis of Variance Table
+
+ Response: leak
+      Df Sum Sq Mean Sq F value    Pr(>F)
+ potato    1 3420.20  3420.20 33.8037 1.6035e-07 ***
+ regime    1 1951.75  1951.75 19.2902 3.8429e-05 ***
+ temp      1 1255.58  1255.58 12.4096 0.00075134 ***
+ Residuals 71 7183.67   101.18
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> print(anova(lm(leak~regime+potato+temp,data=Potato)),digits=6)

+ Analysis of Variance Table
+
+ Response: leak
+      Df Sum Sq Mean Sq F value    Pr(>F)
+ regime    1 3897.00  3897.00 38.5161 3.2325e-08 ***
+ potato    1 1474.96  1474.96 14.5778 0.00028487 ***
+ temp      1 1255.58  1255.58 12.4096 0.00075134 ***
+ Residuals 71 7183.67   101.18
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> print(anova(lm(leak~temp+regime+potato,data=Potato)),digits=6)

+ Analysis of Variance Table
+
+ Response: leak
+
+   Df Sum Sq Mean Sq F value    Pr(>F)
+ temp      1 1171.97  1171.97  11.5832 0.00109736 **
+ regime     1 3952.60  3952.60  39.0656 2.6948e-08 ***
+ potato     1 1502.97  1502.97  14.8546 0.00025228 ***
+ Residuals 71 7183.67   101.18
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results differ due to the way that the sums of squares are computed. In the main effects model, we do have the decomposition

$$SS = SSR_1 + SSR_2 + SSR_3 + SSE$$

but the SSR_j terms can only be associated with the order in which terms are introduced into the model. That is, the sum

$$SSR = SSR_1 + SSR_2 + SSR_3$$

is always the same (see the R-squared statistic and the F-statistic from the initial analyses), but there is no unique way to associate any definitive sum of squares to any individual term.