Modeller vs designer

Modeller vs designer

Let's consider a linear regression with a simple explanatory variable:

$$Y_i = \alpha + \beta_1 x_i + \epsilon_i$$

where

$$\epsilon_i \sim \text{Normal}(0, \sigma^2).$$

Here for observation i

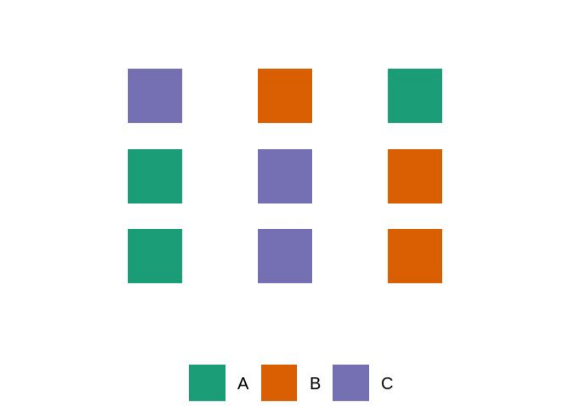
- Y_i is the value of the response
- x_i is the value of the explanatory variable
- ϵ_i is the error term: the difference between Y_i and its expected value
- α is the intercept term (a parameter to be estimated), and
- β_1 is the slope: coefficient of the explanatory variable (a parameter to be estimated)

$$Y_{ik} = \alpha + \tau_k + \epsilon_{ik}$$

where τ_k is called an *effect* and represents the difference between the overall average, α, and the average at the k_{th} treatment level. The errors ϵ_{ik} are again assumed to be normally distributed and independent due to the randomisation (i.e., $\epsilon_{ik} \sim N(0, \sigma^2)$.

Or you might think of the model as

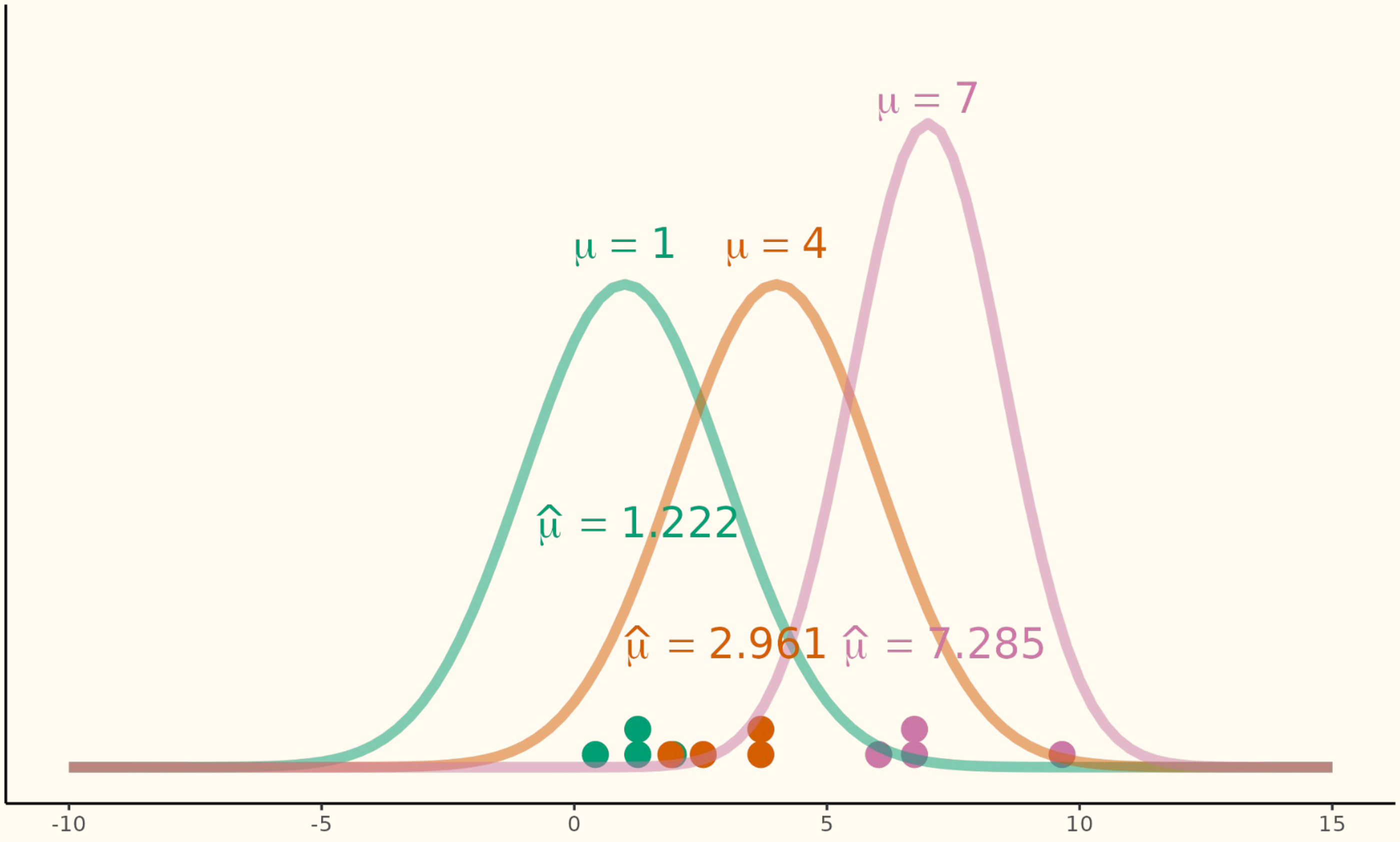
$$Y_{ik} = \mu_k + \epsilon_{ik}$$



Data

Treatment	Response
A	1.95, 1.01, 0.42. 1.45
B	3.79, 2.55, 3.58, 1.91
C	6.56, 6.02, 9.65, 6.90

Data

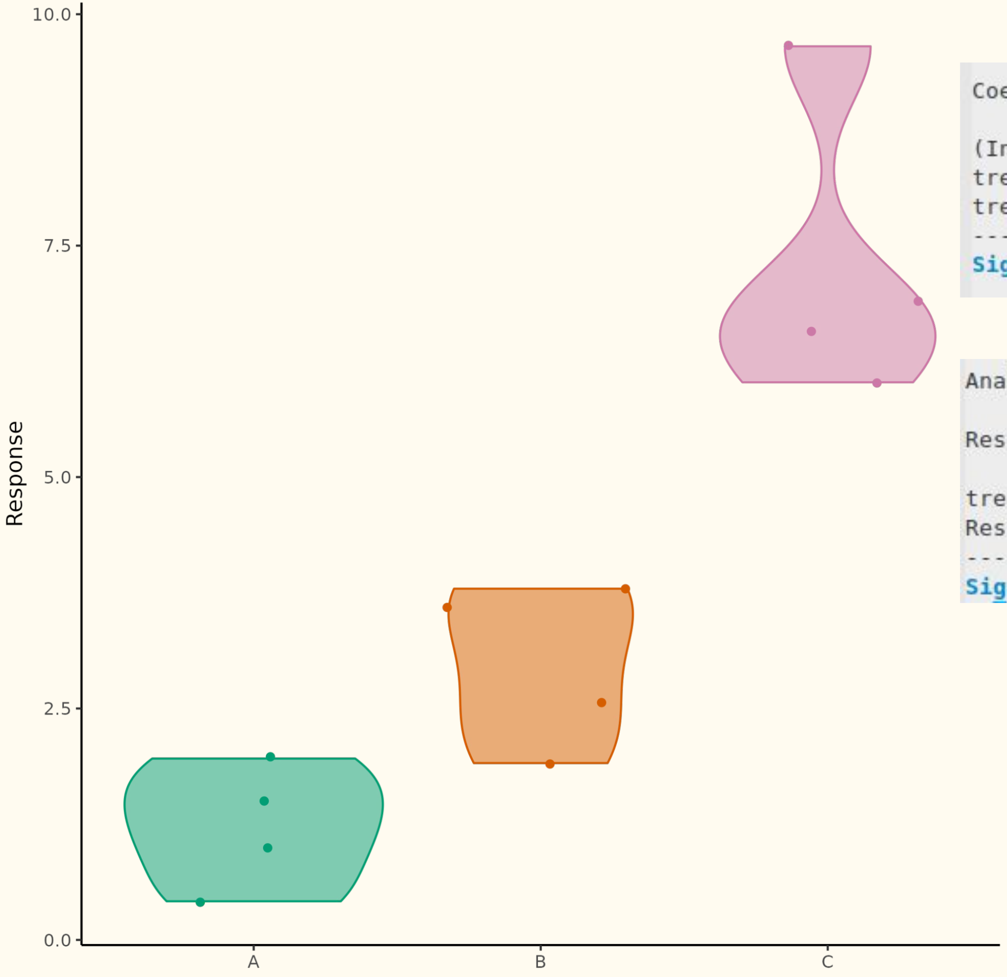


Modeller

A linear model

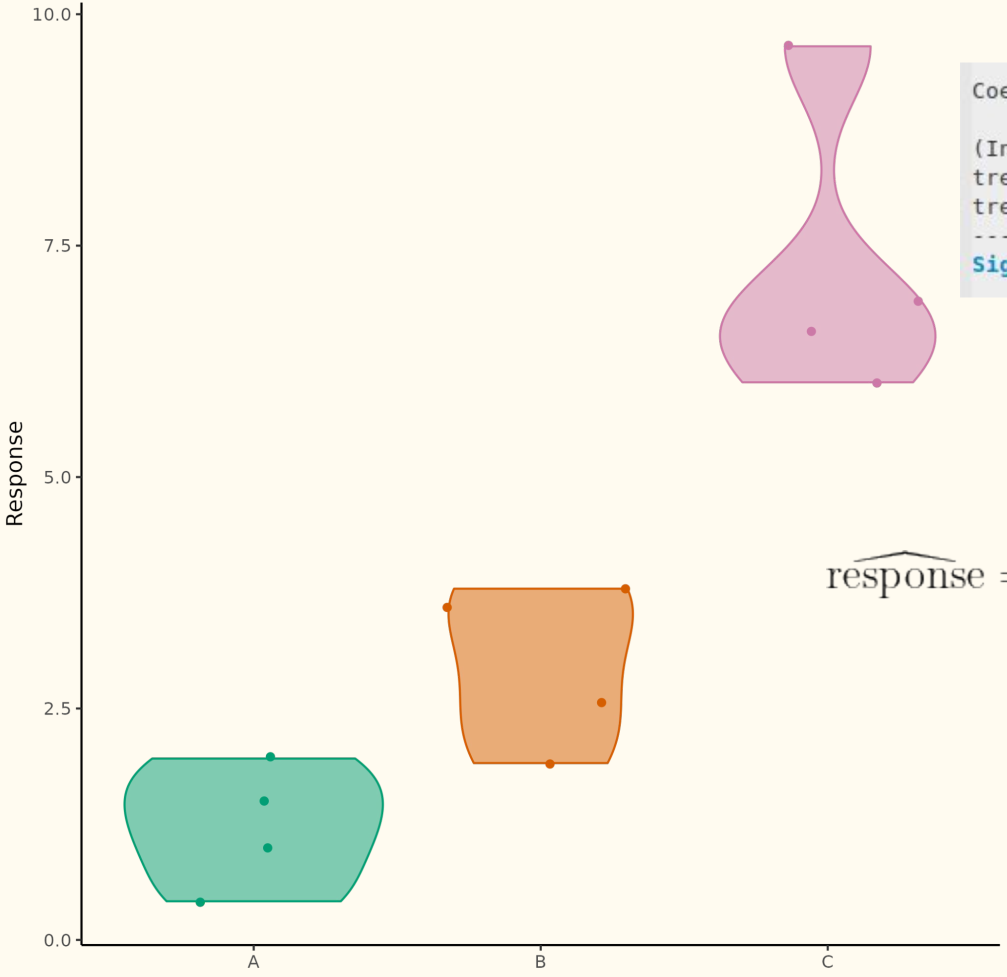
Treatment	Response
A	1.95, 1.01, 0.42. 1.45
B	3.79, 2.55, 3.58, 1.91
C	6.56, 6.02, 9.65, 6.90

$$\text{response} = \alpha + \beta_1(\text{treatment}_\text{B}) + \beta_2(\text{treatment}_\text{C}) + \epsilon$$




```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.2225     0.5661   2.159   0.0591 .
treatmentB    1.7386     0.8006   2.172   0.0580 .
treatmentC    6.0628     0.8006   7.573 3.42e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

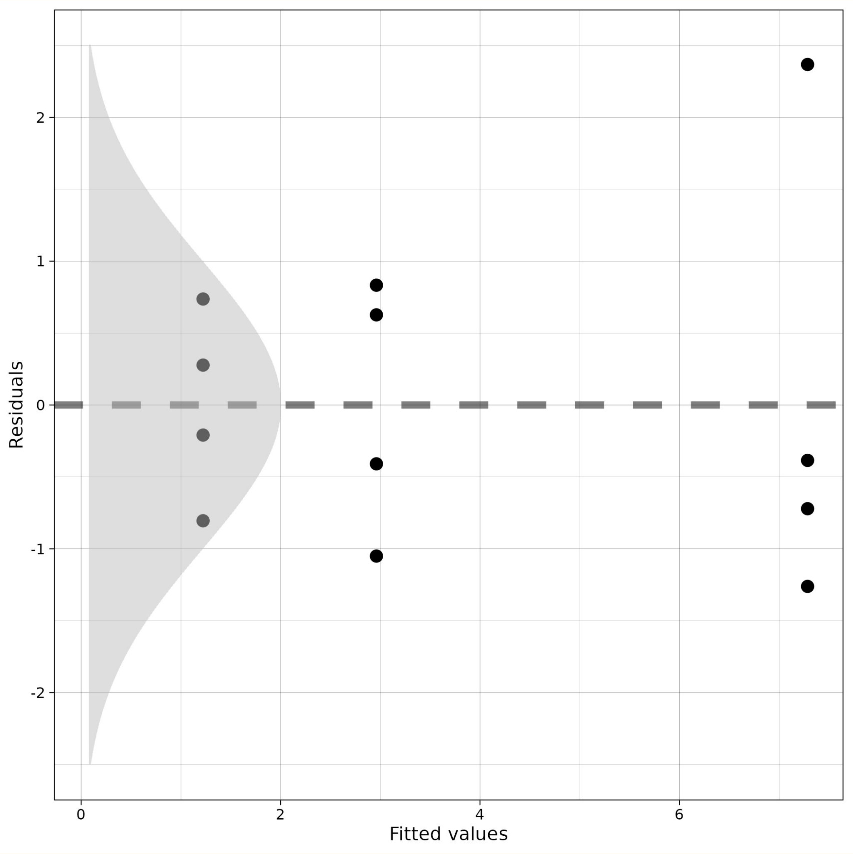
```
Analysis of Variance Table

Response: response
          Df Sum Sq Mean Sq F value    Pr(>F)
treatment  2 77.971  38.985  30.413 9.909e-05 ***
Residuals  9 11.537   1.282
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```




```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.2225     0.5661   2.159   0.0591 .
treatmentB    1.7386     0.8006   2.172   0.0580 .
treatmentC    6.0628     0.8006   7.573 3.42e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$\widehat{\text{response}} = 1.22 + 1.74(\text{treatment}_{\text{B}}) + 6.06(\text{treatment}_{\text{C}})$$

$$\epsilon \sim \mathtt{Normal}(0, \sigma^2)$$



Designer

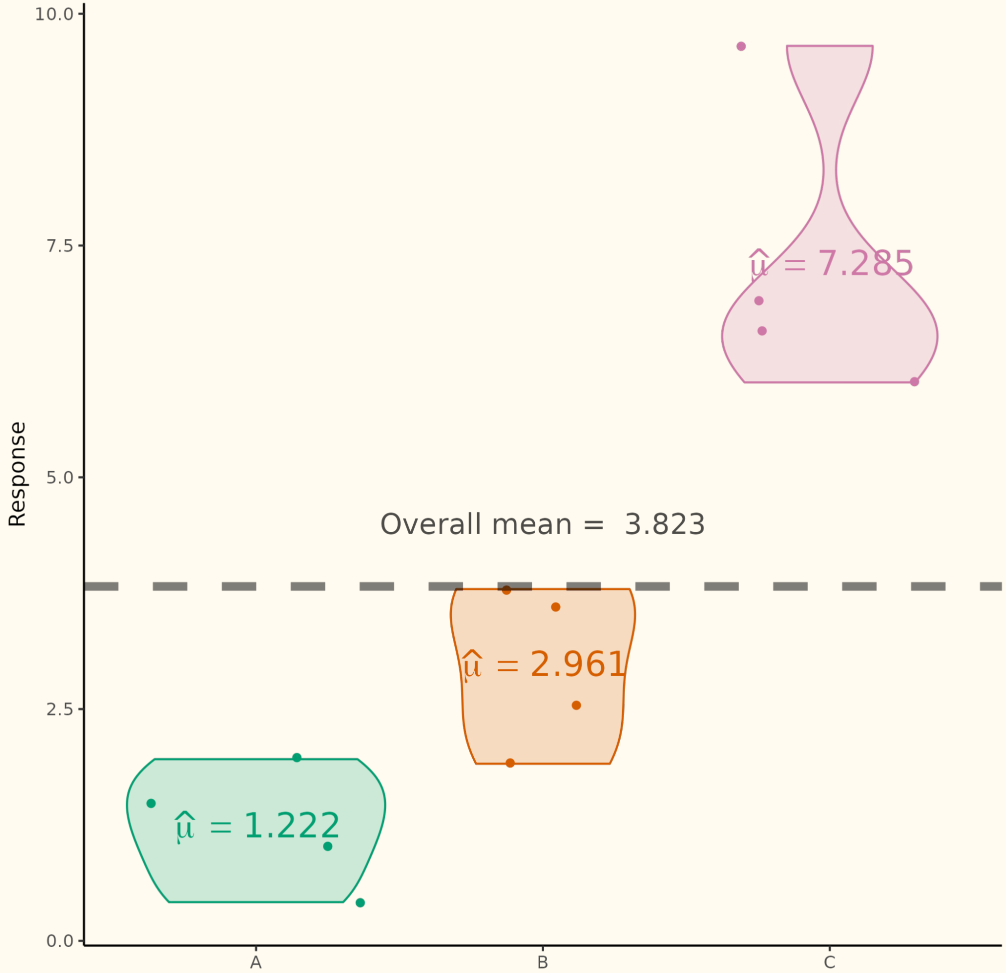
Still, a linear model

Treatment	Response
A	1.95, 1.01, 0.42, 1.45
B	3.79, 2.55, 3.58, 1.91
C	6.56, 6.02, 9.65, 6.90

$$Y_{ij} = \mu_i + \epsilon_{ij} \longrightarrow$$

(for the jth experimental unit subject to the ith level of the treatment factor)

$$Y_{\text{treatment}_A j} = \mu_{\text{treatment}_A} + \epsilon_{\text{treatment}_A j}$$
$$Y_{\text{treatment}_B j} = \mu_{\text{treatment}_B} + \epsilon_{\text{treatment}_B j}$$
$$Y_{\text{treatment}_C j} = \mu_{\text{treatment}_C} + \epsilon_{\text{treatment}_C j}$$



Response
10.0
7.5
5.0
2.5
0.0
A
B
C
Overall mean = 3.823
μ̂ = 7.285
μ̂ = 2.961
μ̂ = 1.222

Measuring distance

$7.285 - 3.823 = 3.462$

$2.961 - 3.823 = -0.862$ $\longrightarrow$ $3.462 - 0.862 - 2.601 = 0$

$1.222 - 3.823 = -2.601$

😱😱😱😱😱😱😱

Measuring distance

(7.285 - 3.823)^2 = 11.98

(2.961 - 3.823)^2 = 0.743 → 19.488

(1.222 - 3.823)^2 = 6.765

4 observations in each group → 4 x 19.488 = 77.952

Treatment	Response	Treatment mean	Overall mean	$\Sigma_{j=1}^{4}(y_j - \mu_{\text{treatment}})^2$
A	1.95, 1.01, 0.42. 1.45	1.22	3.82	1.31
B	3.79, 2.55, 3.58, 1.91	2.96		2.36
C	6.56, 6.02, 9.65, 6.90	7.29		7.87

$$\Sigma_{i=1}^{3}\Sigma_{j=1}^{4}(y_{ij} - \mu_{\text{treatment}})^2 = 1.31 + 2.36 + 7.87 = 11.54$$

$$\Sigma_{i=1}^{3}\Sigma_{j=1}^{4}(y_{ij} - \bar{\mu})^2 = 89.51$$

$$\Sigma_{i=1}^{3}\Sigma_{j=1}^{4}(\mu_{\text{treatment}} - \bar{\mu})^2 = 77.97$$

Actually, we've been a bit lax with notation...

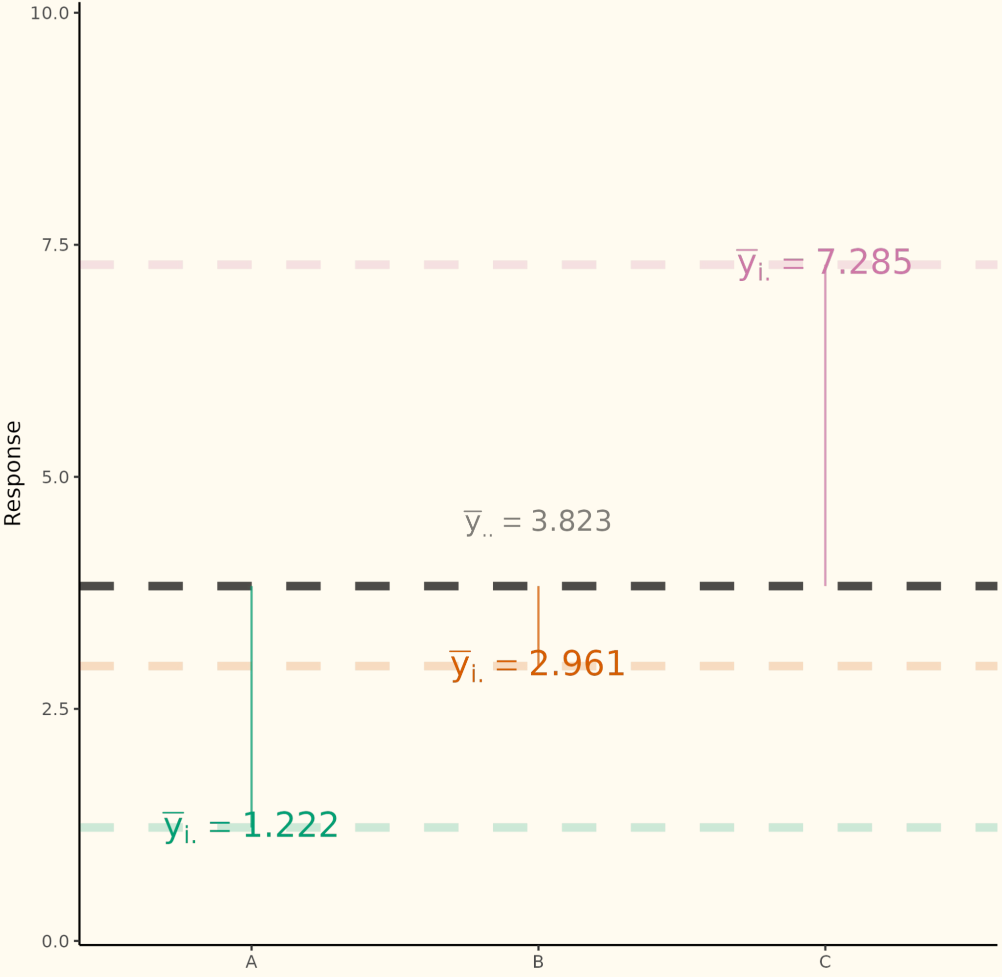
$$SS_{\text{error}} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$

$$SS_{\text{total}} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} y_{ij}^2 - n\bar{y}_{..}^2$$

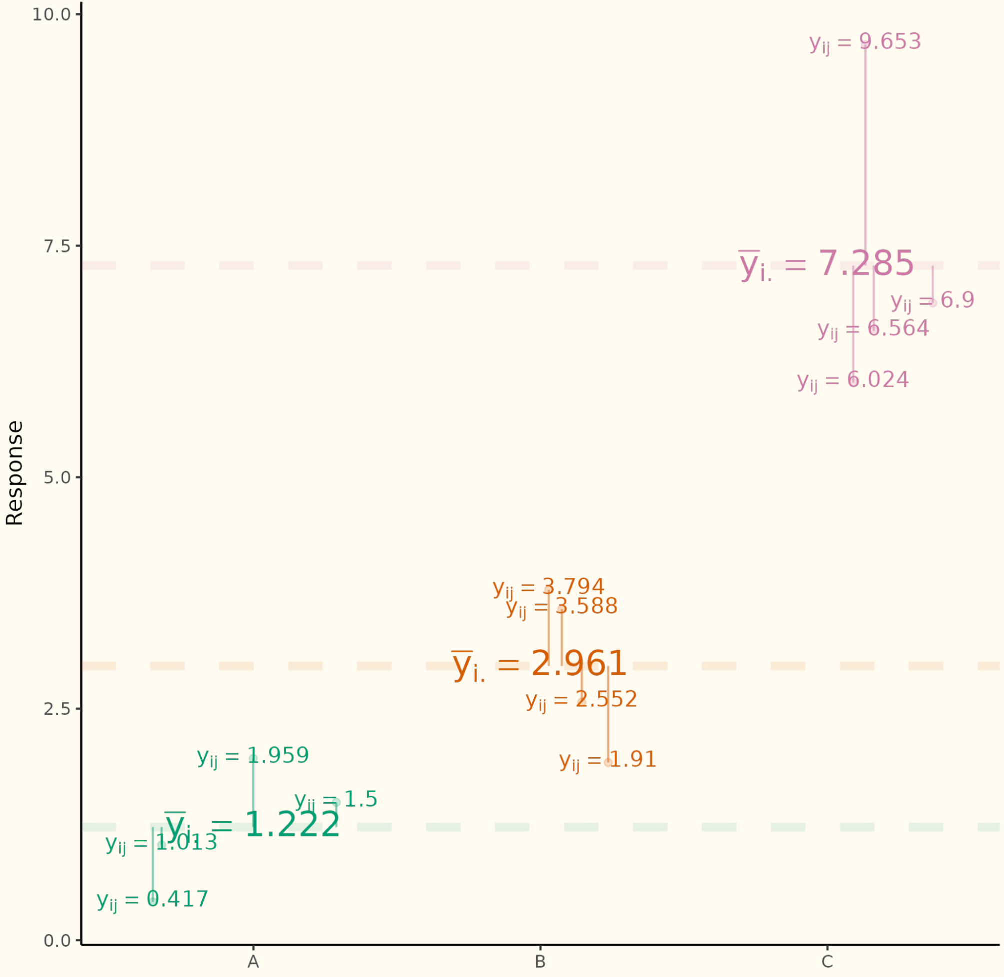
$$SS_{\text{treatment}} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..})^2$$

$$SS_{\text{total}} = SS_{\text{treatment}} + SS_{\text{error}}$$

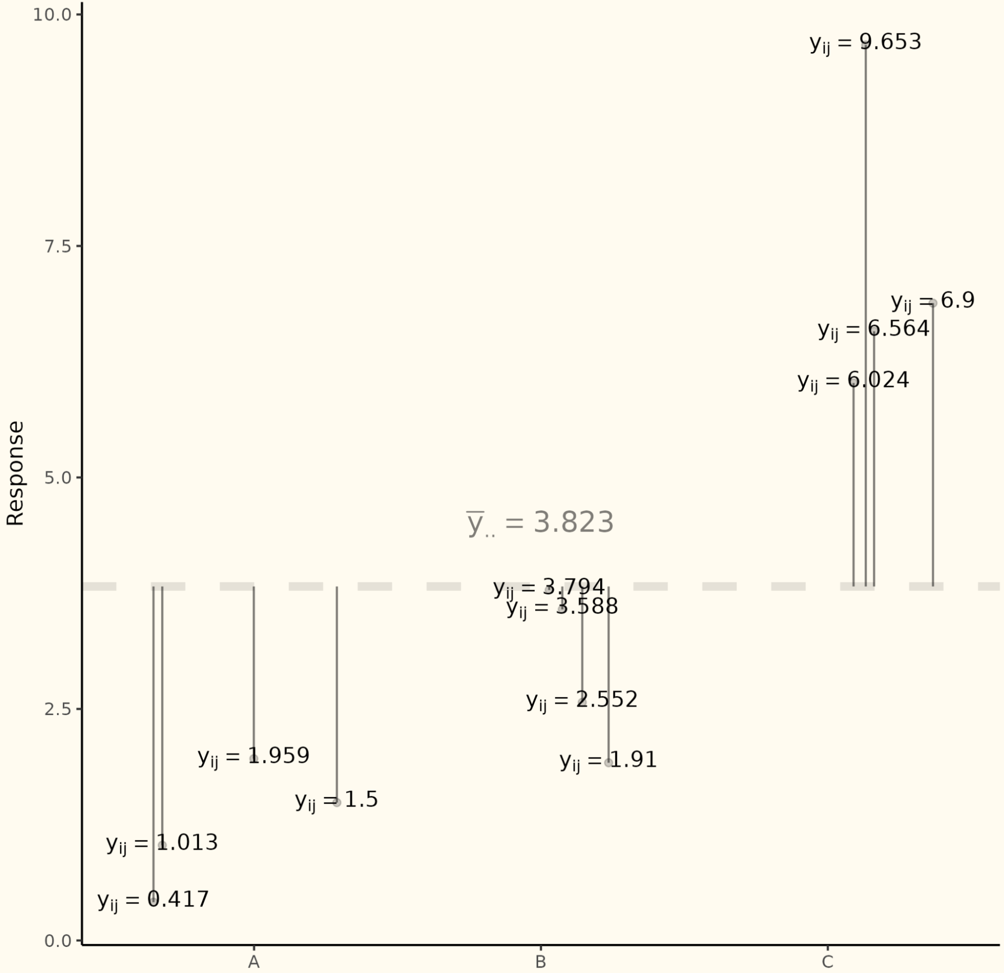
$SS_{\text{treatment}}$



SS_{error}



SS_{total}



Back to the start we go

```
Analysis of Variance Table

Response: response
          Df Sum Sq Mean Sq F value    Pr(>F)    
treatment  2 77.971  38.985  30.413 9.909e-05 ***
Residuals  9 11.537   1.282                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

What if we have two treatments?

The same distance measure is used, however...

Sequential (Type I SS)

- As a term **enters the model** its SS is calculated, which is then **subtracted** from the total SS.
- This then **reduces the available** SS for the next term entering the model.

	W	T
A	*1.95, 0.42*	1.01, 1.45
B	*3.79, 3.58*	2.55, 1.91
C	*6.56, 9.65*	6.02, 6.90

		treatment2	
		W	T
treatment	A	1.95, 0.42	1.01, 1.45
	B	3.79, 3.58	2.55, 1.91
	C	6.56, 9.65	6.02, 6.90

response ~ treatment + treatment2

```
           Df Sum Sq Mean Sq F value   Pr(>F)    
treatment   2  77.97   38.99   36.87 9.18e-05 ***
treatment2  1   3.08    3.08    2.91    0.126    
Residuals   8   8.46    1.06                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

response ~ treatment2 + treatment

```
           Df Sum Sq Mean Sq F value   Pr(>F)    
treatment2  1   3.08    3.08    2.91    0.126    
treatment   2  77.97   38.99   36.87 9.18e-05 ***
Residuals   8   8.46    1.06                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

		treatment2	
		W	T
treatment	A	1.95, 0.42	1.01, 1.45
	B	3.79, 3.58	2.55, 1.91
	C	6.56, 9.65	6.02, 6.90

```
response ~ treatment * treatment2
```

```
                     Df Sum Sq Mean Sq F value   Pr(>F)    
treatment             2  77.97   38.99  34.966 0.000493 ***
treatment2            1   3.08    3.08   2.760 0.147721    
treatment:treatment2  2   1.77    0.89   0.794 0.494435    
Residuals             6   6.69    1.11                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
response ~ treatment2 * treatment
```

```
                     Df Sum Sq Mean Sq F value   Pr(>F)    
treatment2            1   3.08    3.08   2.760 0.147721    
treatment             2  77.97   38.99  34.966 0.000493 ***
treatment2:treatment  2   1.77    0.89   0.794 0.494435    
Residuals             6   6.69    1.11                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What if we have two treatments and our groups are unequal in size?

The same distance measure is used, however…

treatment2

treatment

	W	T
A	1.95, 0.42	1.01, 1.45
B	3.79, 3.58	2.55, 1.91
C	6.56, 9.65	6.02[redacted]

response ~ treatment + treatment2

```
             Df Sum Sq Mean Sq F value  Pr(>F)
treatment     2  67.84   33.92  28.139 0.00045 ***
treatment2    1   2.90    2.90   2.407 0.16474
Residuals     7   8.44    1.21
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

response ~ treatment2 + treatment

```
            Df Sum Sq Mean Sq F value   Pr(>F)
treatment2   1   8.16   8.157   6.767 0.035353 *
treatment    2  62.58  31.292  25.959 0.000578 ***
Residuals    7   8.44   1.205
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

		treatment2	
		W	T
treatment	A	1.95, 0.42	1.01, 1.45
	B	3.79, 3.58	2.55, 1.91
	C	6.56, 9.65	6.02[redacted]

response ~ treatment * treatment2

```
                     Df Sum Sq Mean Sq F value  Pr(>F)
treatment             2  67.84   33.92  26.896 0.00211 **
treatment2            1   2.90    2.90   2.301 0.18977
treatment:treatment2  2   2.13    1.07   0.845 0.48277
Residuals             5   6.31    1.26
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

response ~ treatment2 * treatment

```
                     Df Sum Sq Mean Sq F value  Pr(>F)
treatment2            1   8.16   8.157   6.468 0.05169 .
treatment             2  62.58  31.292  24.812 0.00253 **
treatment2:treatment  2   2.13   1.066   0.845 0.48277
Residuals             5   6.31   1.261
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Sequential (Type I SS)

- As a term **enters the model** its SS is calculated, which is then **subtracted** from the total SS.
- This then **reduces the available** SS for the next term entering the model.
- So… when treatment combinations in a factorial experiment are **unequally replicated**, their effects are **not mutually independent**, so that **the order in which terms enter the model matters.**

Type II SS

- Rather than calculating SS sequentially we can calculate the SS for a given effect **adjusting** for all other effects listed in the model. This means that the SS[A] and SS[B] main effects will both be adjusted for each other (since neither contains the other), but will not be adjusted for SS[A:B] (since it contains both A and B).
- SS[A:B] will be adjusted for **both** main effects.

In R

Type I SS - aov()

response ~ treatment * treatment2

```
                     Df Sum Sq Mean Sq F value  Pr(>F)
treatment             2  67.84   33.92  26.896 0.00211 **
treatment2            1   2.90    2.90   2.301 0.18977
treatment:treatment2  2   2.13    1.07   0.845 0.48277
Residuals             5   6.31    1.26
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

response ~ treatment2 * treatment

```
                     Df Sum Sq Mean Sq F value  Pr(>F)
treatment2            1   8.16   8.157   6.468 0.05169 .
treatment             2  62.58  31.292  24.812 0.00253 **
treatment2:treatment  2   2.13   1.066   0.845 0.48277
Residuals             5   6.31   1.261
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Type II SS - car::Anova()

```
Anova Table (Type II tests)

Response: response
                     Sum Sq Df F value   Pr(>F)
treatment2            2.901  1  2.3005 0.189775
treatment            62.583  2 24.8123 0.002535 **
treatment2:treatment  2.132  2  0.8454 0.482773
Residuals             6.306  5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```