References:

- Chapter 6 of Saville and Wood, *Statistical Methods: A Geometric Approach*.

We continue from last time.

How much does fertilizer affect yield? We record the two groups:

Treatment	y
1	29.80
1	28.57
1	29.97
2	30.33
2	31.27
2	30.37

One-hot / dummy variable encoding of the treatment levels:

Treatment	Value	x_1	x_2
1	$y_{11} = 29.80$	1	0
1	$y_{12} = 28.57$	1	0
1	$y_{13} = 29.97$	1	0
2	$y_{21} = 30.33$	0	1
2	$y_{22} = 31.27$	0	1
2	$y_{23} = 30.37$	0	1

Treatment	Value	x_1	x_2
1	$y_{11} = 29.80$	1	0
1	$y_{12} = 28.57$	1	0
1	$y_{13} = 29.97$	1	0
2	$y_{21} = 30.33$	0	1
2	$y_{22} = 31.27$	0	1
2	$y_{23} = 30.37$	0	1

Model: We can write the model in a few ways:

$$y = \begin{cases} \mu_1 & \text{if } y \text{ is from treatment 1} \\ \mu_2 & \text{if } y \text{ is from treatment 2} \end{cases}$$

As a homogeneous equation (one-hot encoding):

$$y = \mu_1 x_1 + \mu_2 x_2$$

As an inhomogeneous equation (dummy variable encoding with treatment 2 as reference level):

$$y = (\mu_1 - \mu_2) x_1 + \mu_2$$

As an inhomogeneous equation (dummy variable encoding with treatment 1 as reference level):

$$y = \mu_1 + (\mu_2 - \mu_1) x_2$$

Fitted model:

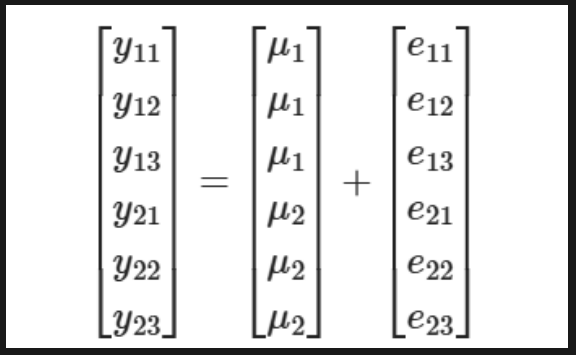
$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_1 \\ \mu_1 \\ \mu_2 \\ \mu_2 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{23} \end{bmatrix}$$

$$\mathrm{df}_{\mathrm{Total}} = ?$$

$$\mathrm{df}_{\mathrm{Model}} = ?$$

$$\mathrm{df}_{\mathrm{Error}} = ?$$

Fitted model:



$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_1 \\ \mu_1 \\ \mu_2 \\ \mu_2 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{23} \end{bmatrix}$$

$$\mathrm{df}_{\mathrm{Total}} = 6.$$

$$\mathrm{df}_{\mathrm{Model}} = 2.$$

$$\mathrm{df}_{\mathrm{Error}} = 6 - 2 = 4.$$

Fitted model (with actual values):

$$\underbrace{\begin{bmatrix} 29.80 \\ 28.57 \\ 29.97 \\ 30.33 \\ 31.27 \\ 30.37 \end{bmatrix}}_{\text{Observation vector}} = \underbrace{\begin{bmatrix} 29.447 \\ 29.447 \\ 29.447 \\ 30.657 \\ 30.657 \\ 30.657 \end{bmatrix}}_{\text{Fitted model vector}} + \underbrace{\begin{bmatrix} .353 \\ -.877 \\ .523 \\ -.327 \\ .613 \\ -.287 \end{bmatrix}}_{\text{Fitted error vector}}$$

Hypothesis testing (preliminary): Our model space $(\mu_1, \mu_1, \mu_1, \mu_2, \mu_2, \mu_2)^T$ is spanned by:

Hypothesis testing (preliminary): Our model space $(\mu_1, \mu_1, \mu_1, \mu_2, \mu_2, \mu_2)^T$ is spanned by:

- the vector $(1, 1, 1, 0, 0, 0)^T$ which tests the hypothesis $H_0 : \mu_1 = 0$, and
- the vector $(0, 0, 0, 1, 1, 1)^T$ which tests the hypothesis $H_0 : \mu_2 = 0$.

Hypothesis testing: Also by

- $(1, 1, 1, -1, -1, -1)^T$ which tests $H_0 : \mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$.
- $(1, 1, 1, 1, 1, 1)^T$ which tests $H_0 : \mu = 0$ where μ is **grand mean** or **overall mean**, estimated by mean $\bar{y}_{\bullet,\bullet}$ of all observations.

Refined orthogonal decomposition:

$$y = \bar{y}_{..} + (\bar{y}_{i\cdot} - \bar{y}_{..}) + (y - \bar{y}_{i\cdot})$$

$$\underbrace{\begin{bmatrix} 29.80 \\ 28.57 \\ 29.97 \\ 30.33 \\ 31.27 \\ 30.37 \end{bmatrix}}_{\text{Observation vector}} = \underbrace{\underbrace{\begin{bmatrix} 30.052 \\ 30.052 \\ 30.052 \\ 30.052 \\ 30.052 \\ 30.052 \end{bmatrix}}_{\text{Overall mean vector}} + \underbrace{\begin{bmatrix} -.605 \\ -.605 \\ -.605 \\ .605 \\ .605 \\ .605 \end{bmatrix}}_{\text{Treatment vector}}}_{\text{Model}} + \underbrace{\begin{bmatrix} .353 \\ -.877 \\ .523 \\ -.327 \\ .613 \\ -.287 \end{bmatrix}}_{\text{Fitted error vector}}$$

ANOVA table:

Source of Variation	df	SS	MS	F
Overall mean ($H_0 : \mu = 0$)	1	5418.6160	5418.6160	
Treatments ($H_0 : \mu_1 = \mu_2$)	1	2.1962	2.1962	5.07 (ns)
Error	4	1.7324	0.4331	
Total	6	5422.5445		

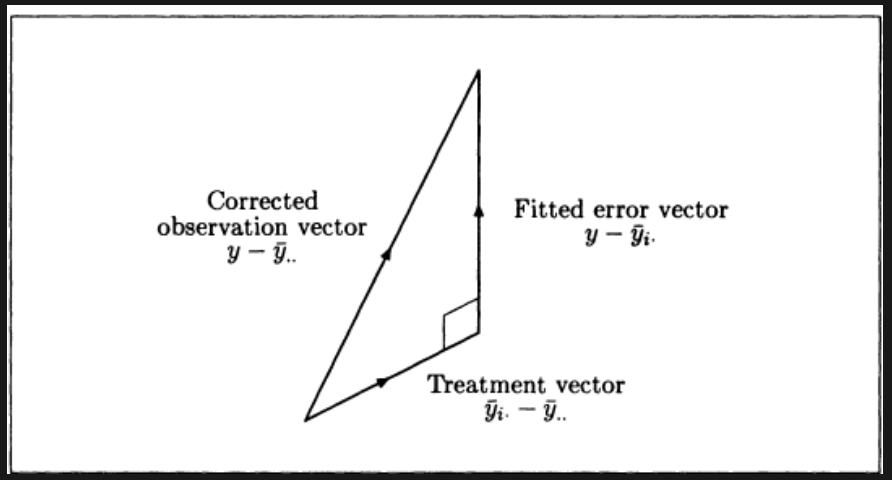
Table 6.2: ANOVA table for our two treatment bulkometer example.

Simplified decomposition:

$$y - \bar{y}_{..} = (\bar{y}_{i\cdot} - \bar{y}_{..}) + (y - \bar{y}_{i\cdot})$$

$$\begin{bmatrix} -.252 \\ -1.482 \\ -.082 \\ .278 \\ 1.218 \\ .318 \end{bmatrix} = \begin{bmatrix} -.605 \\ -.605 \\ -.605 \\ .605 \\ .605 \\ .605 \end{bmatrix} + \begin{bmatrix} .353 \\ -.877 \\ .523 \\ -.327 \\ .613 \\ -.287 \end{bmatrix}$$

Simplified decomposition:



Simplified ANOVA table:

Source of Variation	df	SS	MS	F
Treatments ($H_0 : \mu_1 = \mu_2$)	1	2.1962	2.1962	5.07 (ns)
Error	4	1.7324	.4331	
Total (corrected)	5	3.9286		

Table 6.3: The more usual form of the ANOVA table for the two treatment bulkometer example.

General case:

- Group 1 with n_1-many values $y_{11}, y_{12}, \ldots, y_{1n_1}$;
- Group 2 with n_2-many values $y_{21}, y_{22}, \ldots, y_{2n_2}$;

Then:

Source of Variation	df	SS	MS
Treatments	1	$\|\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}\|^2$	$\|\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}\|^2$
Error	$n_1 + n_2 - 2$	$\|y - \bar{y}_{i\cdot}\|^2$	$\dfrac{\|y - \bar{y}_{i\cdot}\|}{n_1 + n_2 -}$
Total (corrected)	$n_1 + n_2 - 1$	$\|y - \bar{y}_{\cdot\cdot}\|^2$	

Assumptions: This model assumes each group has equal variance. If this assumption fails, then use a more general form of ANOVA called Welch's ANOVA (like Welch's t-test).