## References:

- Chapters 5 and 6 of Saville and Wood, *Statistical Methods: A Geometric Approach*.
- Appendix A of Saville and Wood, *Statistical Methods: A Geometric Primer*.

# SINGLE POPULATION QUESTIONS

How much does fertilizer affect yield? We record the change in yield on three farms:

| Farms | Change (tonnes/ha) |
|---|---|
| 1 | $y_1 = +1.1$ |
| 2 | $y_2 = +1.4$ |
| 3 | $y_3 = +1.1$ |

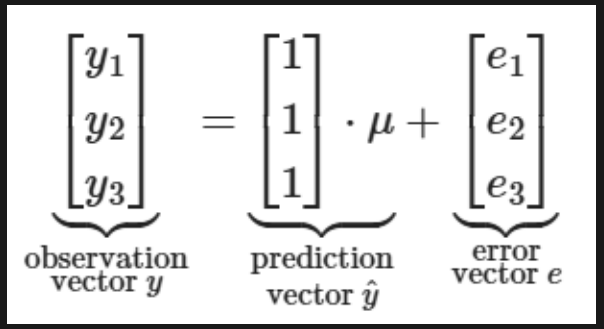
How much does fertilizer affect yield? We record the change in yield on three farms:

| Farms | Change (tonnes/ha) |
|---|---|
| 1 | $y_1 = +1.1$ |
| 2 | $y_2 = +1.4$ |
| 3 | $y_3 = +1.1$ |

**Model:** $y = \mu$

# Fitted model and degrees of freedom:

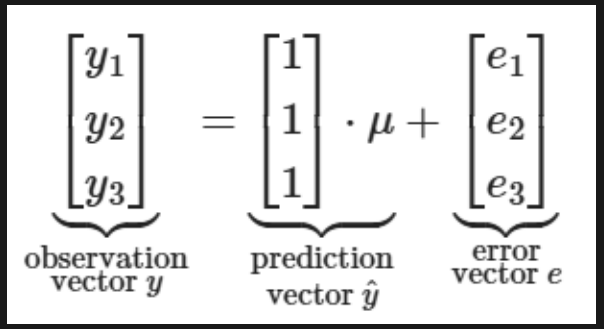
$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}}_{\text{observation vector } y} = \underbrace{\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \cdot \mu}_{\text{prediction vector } \hat{y}} + \underbrace{\begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}}_{\text{error vector } e}$$




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

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

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




$$\mathrm{df}_{\mathrm{Total}} = 3.$$

$$\mathrm{df}_{\mathrm{Model}} = 1.$$

$$\mathrm{df}_{\mathrm{Error}} = 3 - 1 = 2.$$

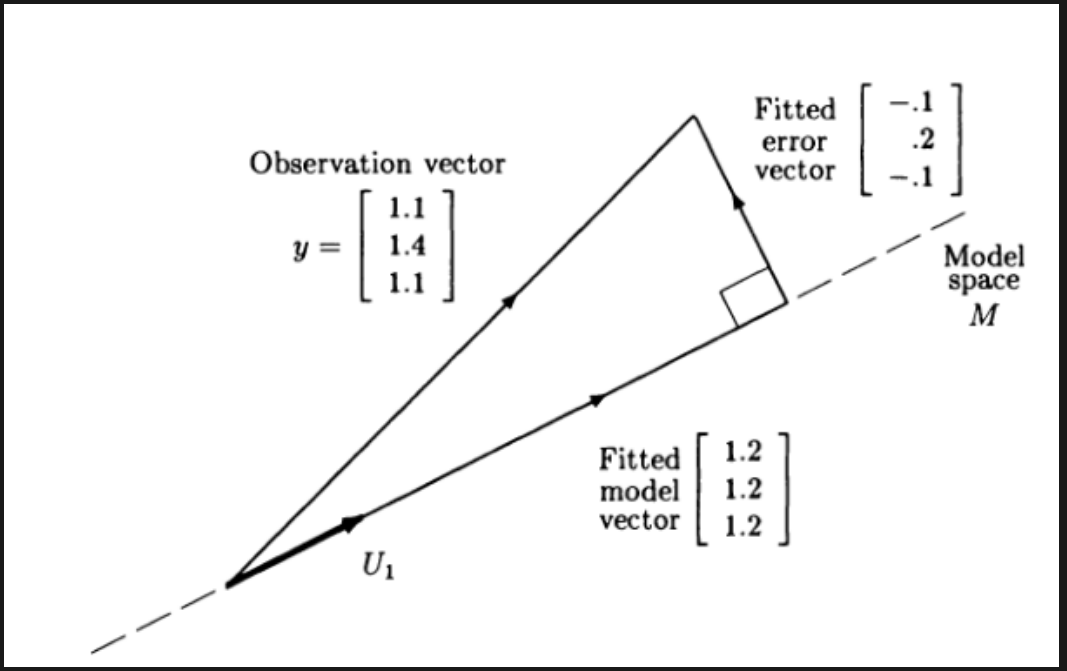
# Fitted model (with actual values):

$$\underbrace{\begin{bmatrix}1.1\\1.4\\1.1\end{bmatrix}}_{\text{observation vector } y} = \underbrace{\begin{bmatrix}1\\1\\1\end{bmatrix} \cdot 1.2}_{\text{prediction vector } \hat{y}} + \underbrace{\begin{bmatrix}-0.1\\+0.2\\-0.1\end{bmatrix}}_{\text{error vector } e}$$

# Fitted model (with actual values):

$$\underbrace{\begin{bmatrix}1.1\\1.4\\1.1\end{bmatrix}}_{\text{observation vector } y} = \underbrace{\begin{bmatrix}1.2\\1.2\\1.2\end{bmatrix}}_{\text{prediction vector } \hat{y}} + \underbrace{\begin{bmatrix}-0.1\\+0.2\\-0.1\end{bmatrix}}_{\text{error vector } e}$$

# Picture:


Observation vector
$y = \begin{bmatrix} 1.1 \\ 1.4 \\ 1.1 \end{bmatrix}$
Fitted error vector $\begin{bmatrix} -.1 \\ .2 \\ -.1 \end{bmatrix}$
Model space $M$
Fitted model vector $\begin{bmatrix} 1.2 \\ 1.2 \\ 1.2 \end{bmatrix}$
$U_1$

# ANOVA table: we explained how to compute df

| Source of Variation | Degrees of freedom (df) | Sum of Squares (SS) | Mean Square (MS) = SS/df | F |
|---|---|---|---|---|
| Model | 1 | | | |
| Error | 2 | | | |
| Total | 3 | | | |

# ANOVA table: SS

| Source of Variation | Degrees of freedom (df) | Sum of Squares (SS) | Mean Square (MS) = SS/df | F |
|---|---|---|---|---|
| Model | 1 | 4.32 | | |
| Error | 2 | 0.06 | | |
| Total | 3 | 4.38 | | |

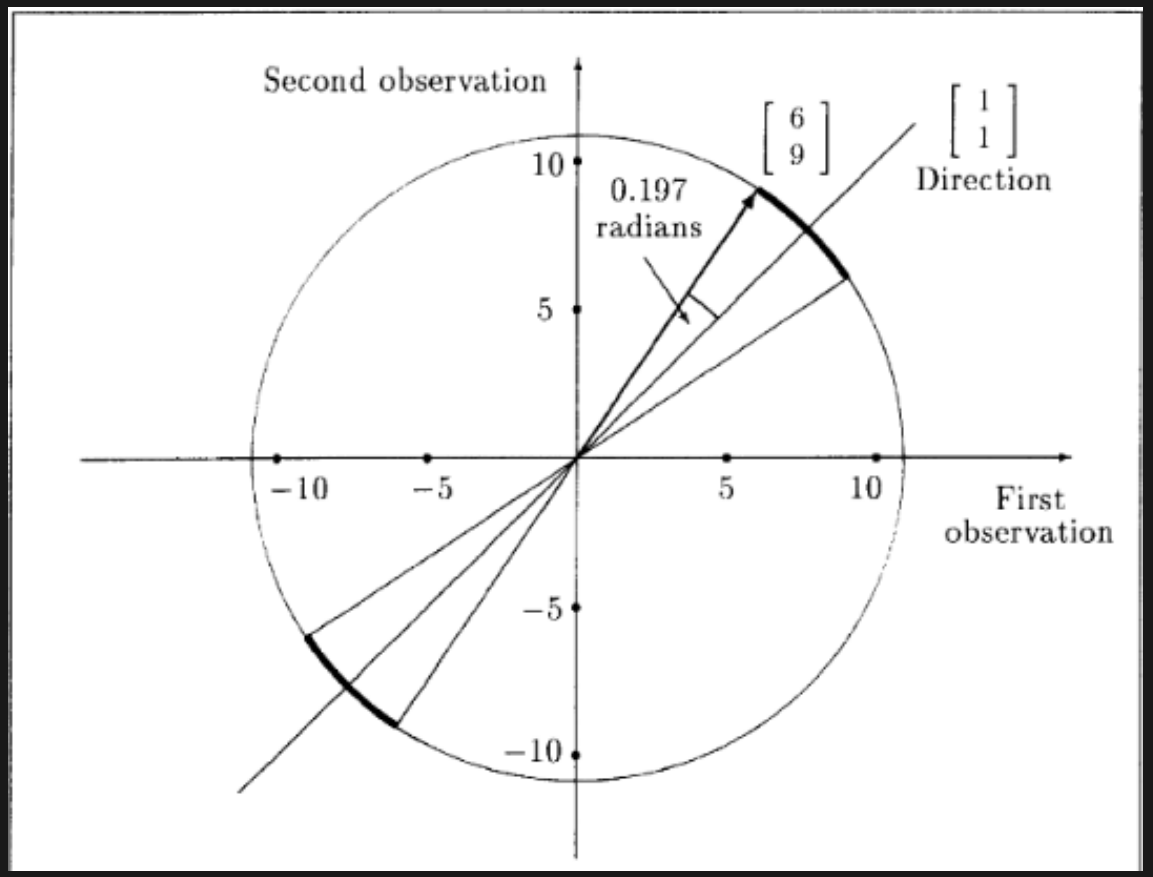
# ANOVA table: MS and F

| Source of Variation | Degrees of freedom (df) | Sum of Squares (SS) | Mean Square (MS) = SS/df | F |
|---|---|---|---|---|
| Model | 1 | 4.32 | 4.32 | 144 |
| Error | 2 | 0.06 | 0.03 | |
| Total | 3 | 4.38 | | |

# ANOVA table: $R^2$ and adjusted $R^2$

| Source of Variation | Degrees of freedom (df) | Sum of Squares (SS) | Mean Square (MS) = SS/df | F |
|---|---|---|---|---|
| Model | 1 | 4.32 | 4.32 | $144$ |
| Error | 2 | 0.06 | 0.03 | |
| Total | 3 | 4.38 | | |

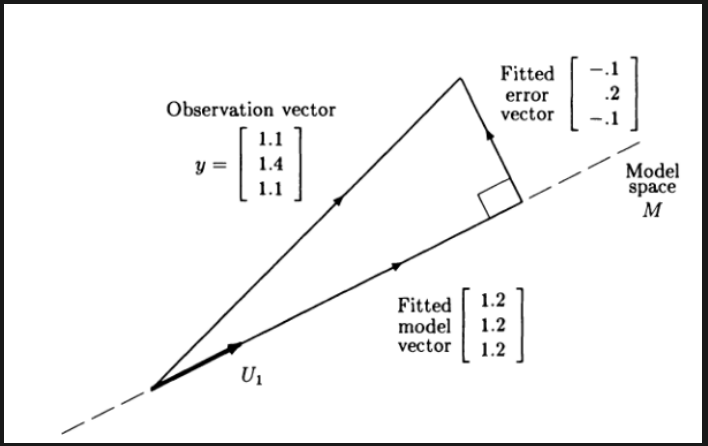
# Geometric interpretation of the $p$-value:



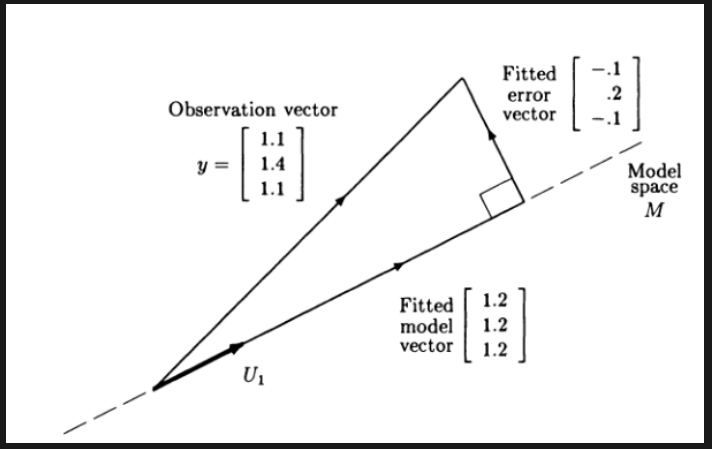
In general,

$$p = \frac{\text{Surface area (caps of cones)}}{\text{Surface area (hypersphere)}} = \frac{\displaystyle\int_0^{\theta} \sin^{n-2} u, du}{\displaystyle\int_0^{\pi/2} \sin^{n-2} u, du}$$

# Geometric interpretation of the $F$-statistics:



# Geometric interpretation of the $F$-statistics:




$$F_{1,df_E} = \frac{\|\bar{y}\|^2}{\|e\|/\mathrm{df}_E} = \mathrm{df}_E \cdot \frac{\|\bar{y}\|^2}{\|e\|^2} = \mathrm{df}_E \cdot \cot^2\theta$$

Work on group problems now.

# QUESTIONS ABOUT TWO POPULATIONS

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:**

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

Or homogeneous model (one-hot encoding):

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

Or inhomogeneous model with $x_1$:

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

Or inhomogeneous model with $x_2$:

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

## Fitted model:

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

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

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

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