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	$y_3 = +1.1$

Model:

$$y = \mu$$

Fitted model and degrees of freedom: Since we have three observations y_1, y_2, y_3, the fitted model should look like:

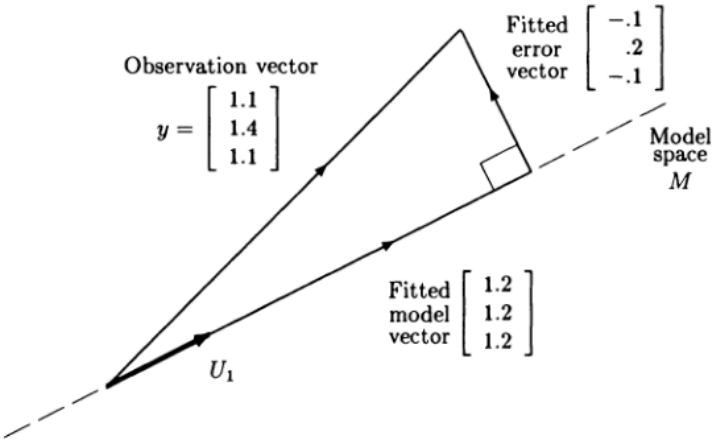
$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}}_{\substack{\text{observation} \\ \text{vector} \\ y}} = \underbrace{\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \cdot \mu}_{\substack{\text{prediction} \\ \text{aka model} \\ \text{vector} \\ \hat{y}}} + \underbrace{\begin{bmatrix} e_1 \\ e_2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

In general, the dimension of the vector space of possible observations is 3, so we say the total degrees of freedom is $\mathrm{df}_{\mathrm{Total}} = 3$.

Our model space is spanned by the vector $(1, 1, 1)^T$ with a point on it specified by the free parameter μ, which defines a subspace of dimension 1, so we write $\mathrm{df}_{\mathrm{Model}} = 1$.

This leaves the error space to have remaining dimension of $\mathrm{df}_{\mathrm{Error}} = 3 - 1 = 2$.

Fitted model (with actual values) from the picture:



The model vector is chosen to be the closest point on the model space to the observation vector. This has two immediate consequences:

- first, the error vector $e = y - \hat{y}$ will be the shortest vector among all choices of $\hat{y}$ within the model space;
- second, e will be orthogonal to the model space and, in particular, the model vector.

Some computation shows that μ should be estimated by the sample mean, so $\hat{\mu} = \bar{y}$:

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

Or:

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

The squared lengths of these vectors are called the **sum of squares total** (**SST**), **sum of squares error** (**SSE**), and **sum of squares model** (**SSM**) in the ANOVA table:

$$SSM = \|\hat{y}\|^2 = \left\| \begin{bmatrix} 1.2 \\ 1.2 \\ 1.2 \end{bmatrix} \right\|^2 = (1.2)^2 + (1.2)^2 + (1.2)^2 = \mathbf{4.32}$$

$$SSE = \|e\|^2 = \left\| \begin{bmatrix} -0.1 \\ 0.2 \\ -0.1 \end{bmatrix} \right\|^2 (-0.1)^2 + (0.2)^2 + (-0.1)^2 = \mathbf{0.06}$$

$$SST = \|y\|^2 = \left\| \begin{bmatrix} 1.1 \\ 1.4 \\ 1.1 \end{bmatrix} \right\|^2 = (1.1)^2 + (1.4)^2 + (1.1)^2 = \mathbf{4.38}$$

Because the vectors $\hat{y}$ and e are orthogonal, the **Pythagorean identity** holds:

$$SST = SSM + SSE$$
$$\mathbf{4.38} = 4.32 + 0.06$$

We have now described the first two numeric columns in the ANOVA table; the **mean square** value is the sum of square divided by the corresponding degree of freedom:

Source of Variation	**Degrees of freedom (df)**	**Sum of Squares (SS)**	**Mean Square (MS) = SS/df**	**F**
Subspace	**Dimension of subspace**	**Sum of squared projection lengths**	**Mean squared length**	**Ratio of mean squares**
Model	1	4.32	4.32	144
Error	2	0.06	0.03	
Total	3	4.38		

Given a sample size of n, here is the general ANOVA table:

Source of Variation	**df**	**SS**	**MS**	**F**
Mean	1	$\|\|\bar{y}\|\|^2$	$\|\|\bar{y}\|\|^2$	$\frac{\|\|\bar{y}\|\|^2}{\|\|y-\bar{y}\|\|^2/(n-1)}$
Error	$n-1$	$\|\|y-\bar{y}\|\|^2$	$\frac{\|\|y-\bar{y}\|\|^2}{n-1}$	
Total	n	$\|\|y\|\|^2$		

Additional statistics:
The statistics R^2, adjusted R^2, and F are defined in terms of the ANOVA table:

$$R^2 = \frac{SSM}{SST}, \qquad R^2_{adj} = \frac{MSM}{MST}, \qquad F_{df_M, df_E} = F_{1,n-1} = \frac{MSM}{MSE}$$

When $df_M = 1$, the SSM consists of the sum of squares of exactly one term, so the sign of its square root is meaningful, and we recover the usual t-value as the square root of F:

$$F_{1,n-1} = \frac{\|\bar{y}\|^2}{\|y-\bar{y}\|^2/(n-1)} = \frac{n\bar{y}^2}{s^2} = \frac{\bar{y}^2}{(s^2/n)} = \left(\frac{\bar{y}}{s/\sqrt{n}}\right)^2 = t^2$$

Here s is the sample standard deviation.

Geometric interpretation of the p-value:
The p-value is the chance your observation vector makes an angle as close as, or closer than, the model vector $(1, \ldots, 1)^T$. In 2-dimensions, the p-value is therefore the lengths of the shaded arc divided by the total circumference of the circle:

$$p = \frac{4\theta}{2\pi}$$

From page 184 of Geometric Primer:

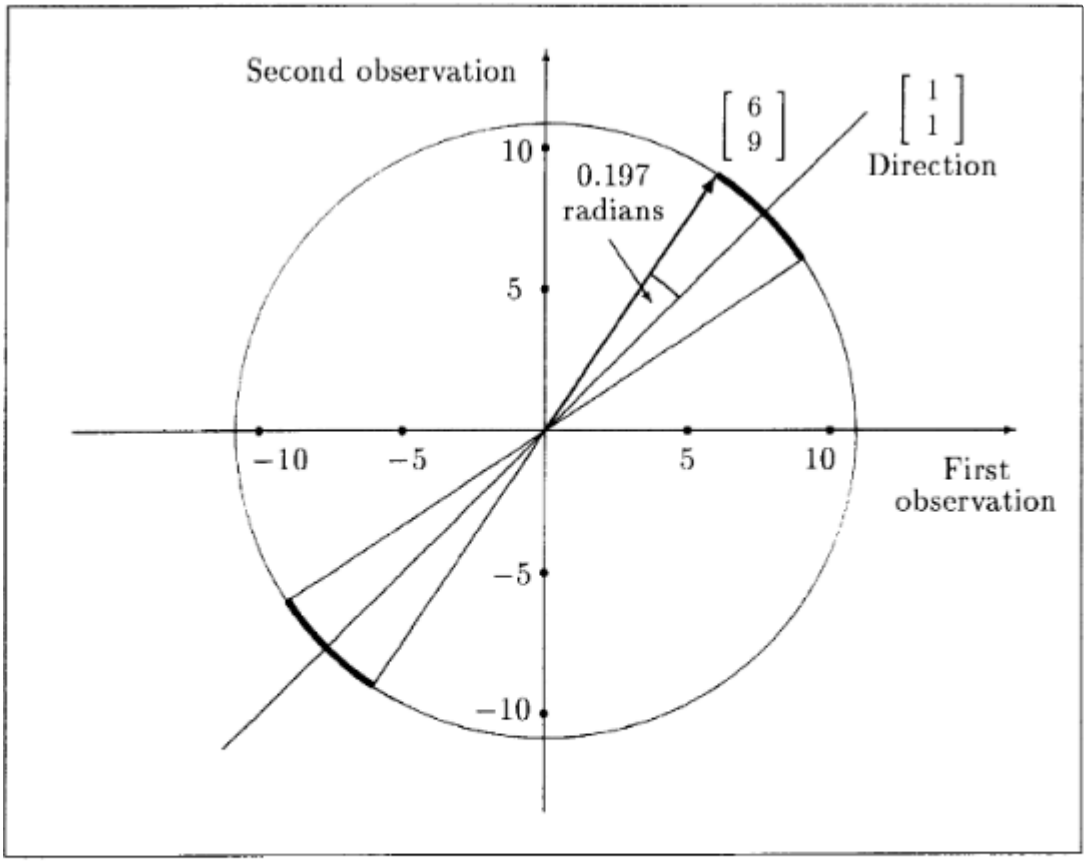


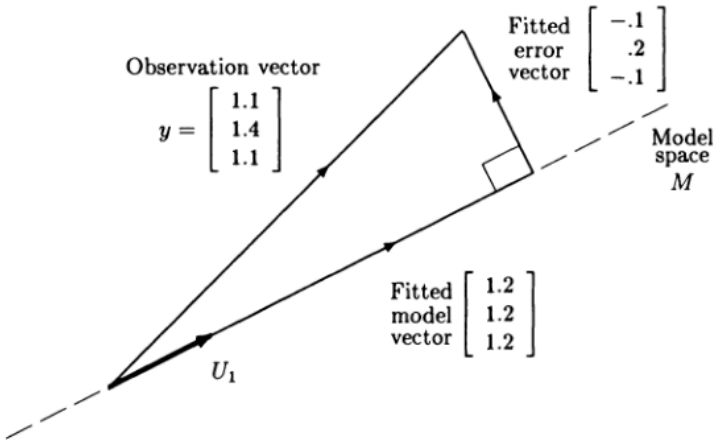
Figure D.4: Our vector of observations $[6, 9]^{\mathrm{T}}$ makes an angle of 0.197 radians with the equiangular direction. If $\mu = 0$, what is the probability of observing an angle as small, or smaller than this? The answer is the ratio of the lengths of the highlighted arcs to the circumference of the circle, $p = 4 \times 0.197/(2\pi) = 0.197/(\pi/2) = 0.13$.

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:

Look back at the picture.



Note that $\|y\|$ and $\|e\|$ are the side lengths of a right triangle which relate to the angle θ of this triangle via the cotangent function. Hence:

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

The intermediate statistics θ allows us to convert between the F-statistics and the p-value.

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

We can also write this with one-hot encoding or 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

Treatment	Value	x_1	x_2
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):

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

Using only x_1 gives a different equation, with treatment 2 as reference level:

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

Using only x_1 gives a different equation, with treatment 1 as reference level:

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

In all cases, μ_i is estimated as the sample mean of that group i:

$$\hat{\mu}_i = \bar{y}_{i\bullet} = \frac{y_{i1} + y_{i2} + y_{i3}}{3}$$

⚠ SAS and R drop redundant variables when exact collinearity occurs

Since $x_2 = 1 - x_1$, the variables x_1 and x_2 are linearly dependent (exact collinearity), so `PROC REG` in SAS and `lm()` in R will automatically drop the redundant variable x_2 from the model by assigning its value as `<NA>` if we use x_1, x_2, and allow an intercept:

```
Call:
lm(formula = y ~ x_1 + x_2, data = data)

Coefficients:
(Intercept)          x_1          x_2
      30.66        -1.21           NA
```

If an intercept is not included, `R` happily uses both dummy variables:

```
Call:
lm(formula = y ~ x_1 + x_2 - 1, data = data)

Coefficients:
  x_1    x_2
29.45  30.66
```

(Lesson to be continued next time.)