Vector Geometry for Statistical Models

Andrew Zieffler

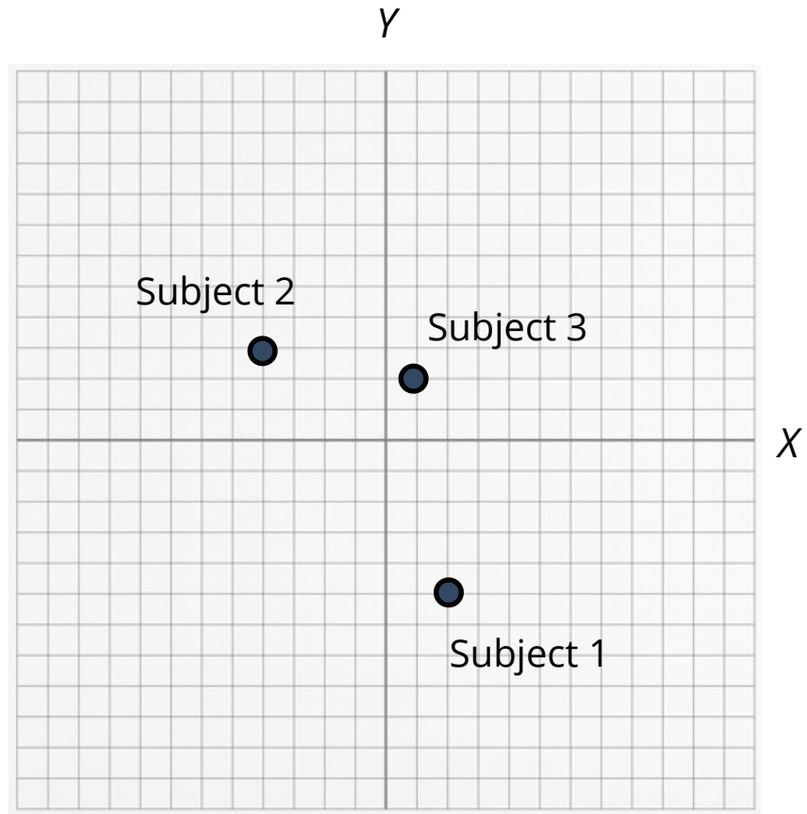
Statistical Geometry

It is possible to compute all statistical calculations for regression models using only a ruler and protractor. The point here is not to displace the computer (it can do this faster), but, rather to help you understand some of the concepts at a deeper level.

Variable Space

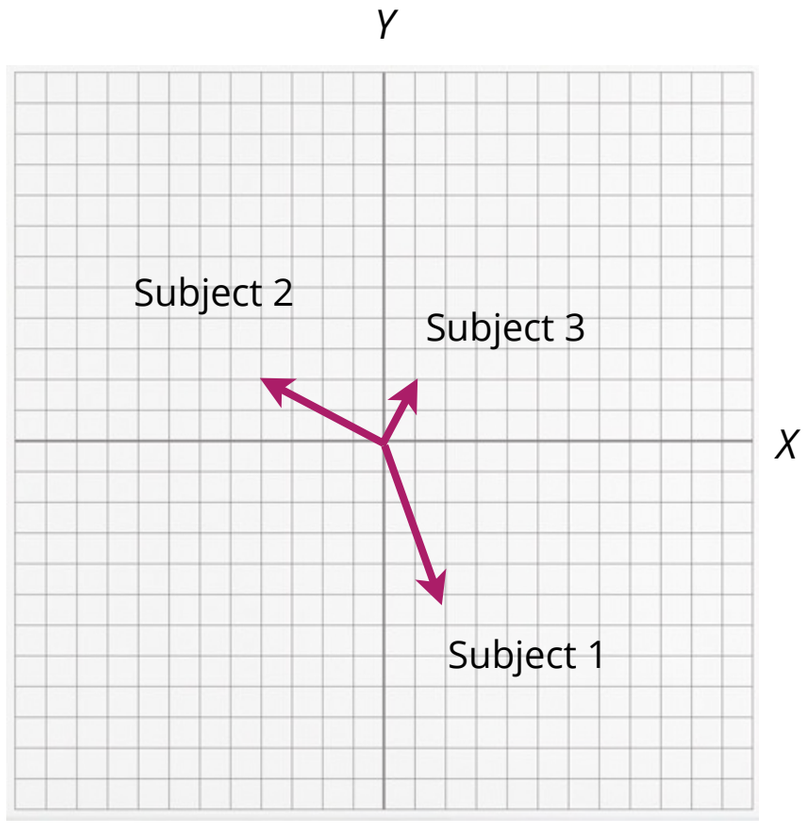
We typically plot multivariate data in variable space. In variable space, the **variables are represented as axes** and the **subjects are represented as points**, which are plotted based on their values for the variables.

Subject	**X**	**Y**
Subject 1	2	–5
Subject 2	–4	3
Subject 3	1	2



Rather than points, we could also draw vectors. Representing the subjects' values on the variables with a point or vector is just a matter of convenience.

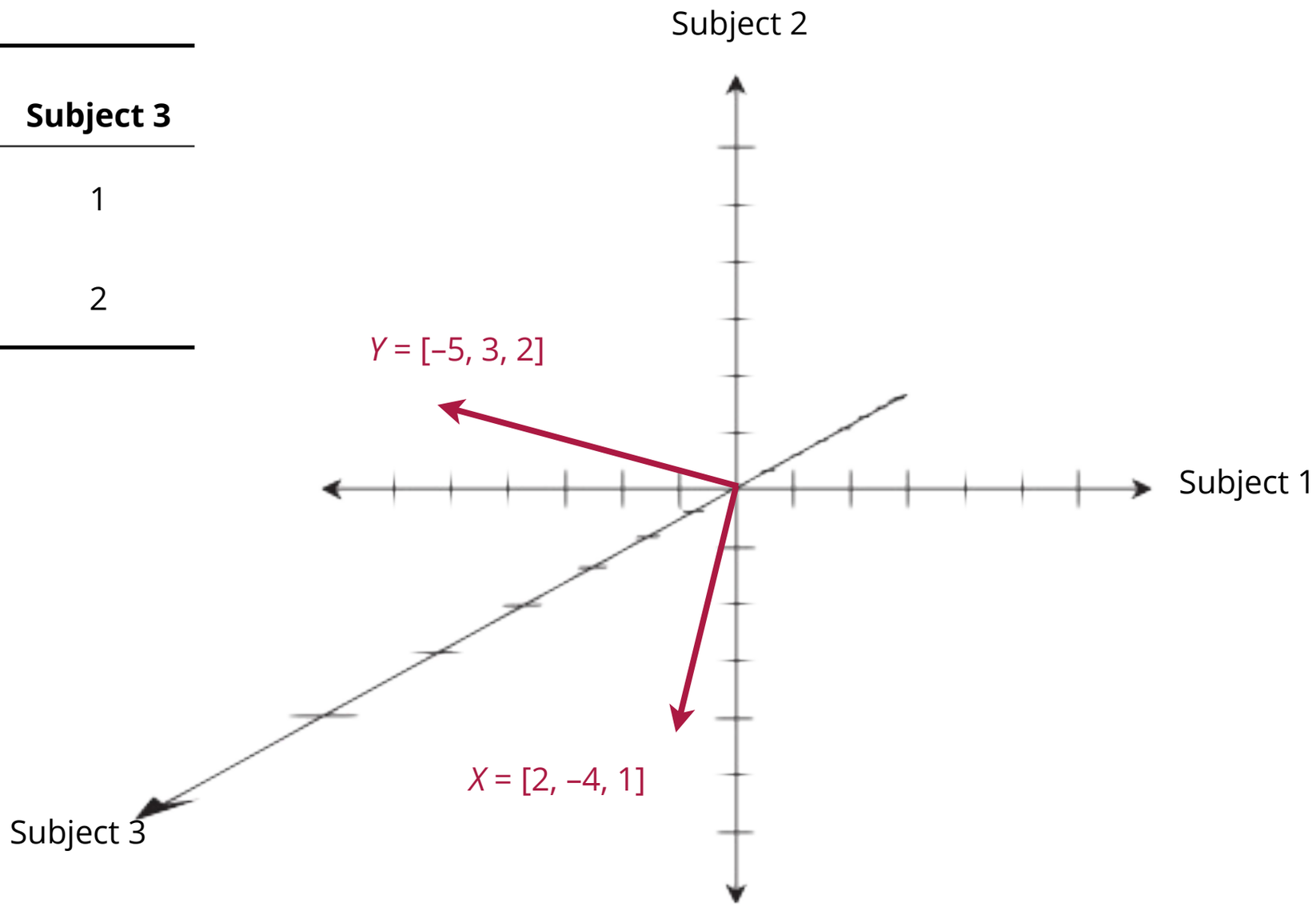
Subject	X	Y
Subject 1	2	–5
Subject 2	–4	3
Subject 3	1	2



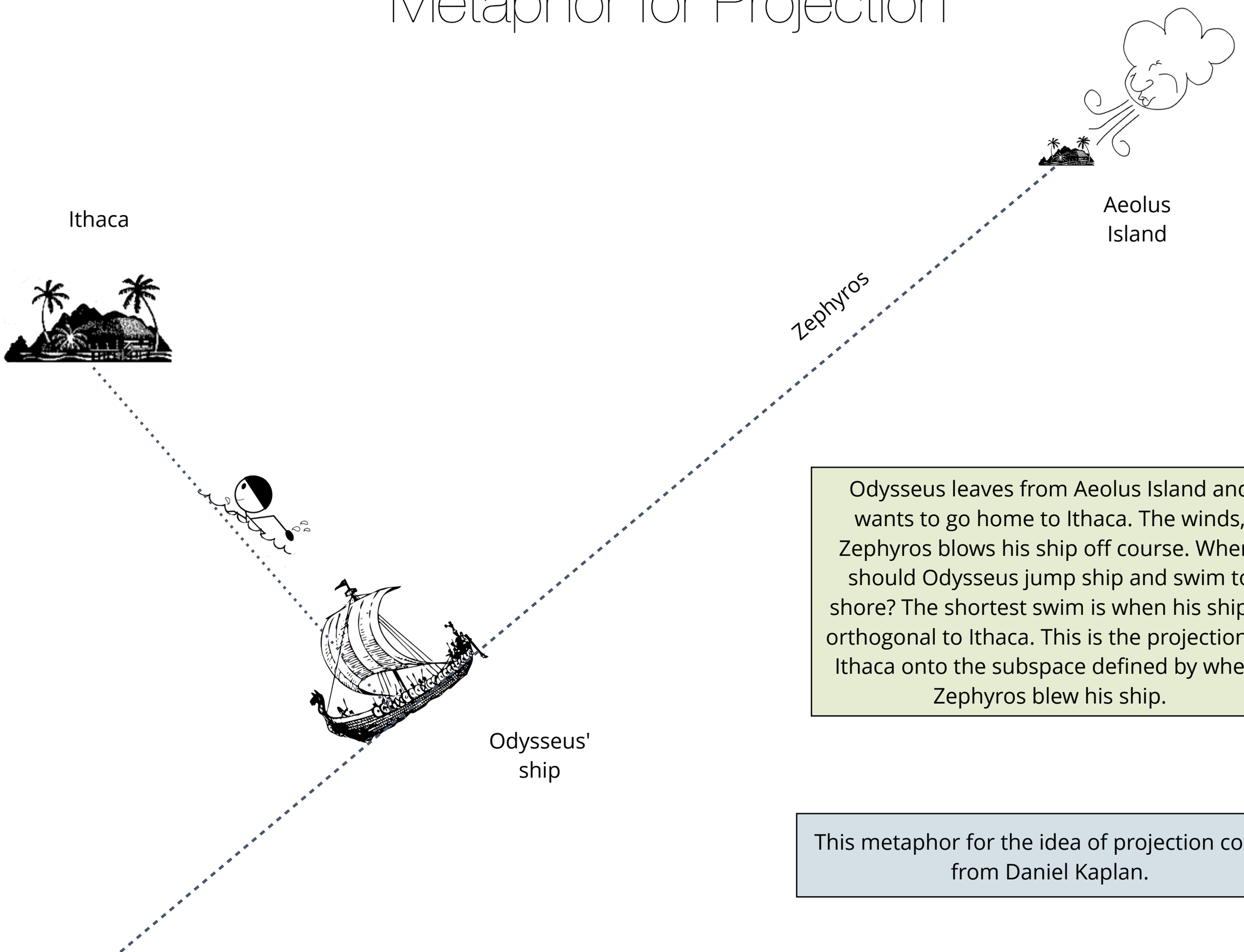
Subject Space

In subject space, the **subjects are represented as axes** and the **variables are represented as vectors**. The value of subject space is not in plotting, but rather in understanding statistical modeling.

Variable	Subject 1	Subject 2	Subject 3
X	2	–4	1
Y	–5	3	2



Metaphor for Projection



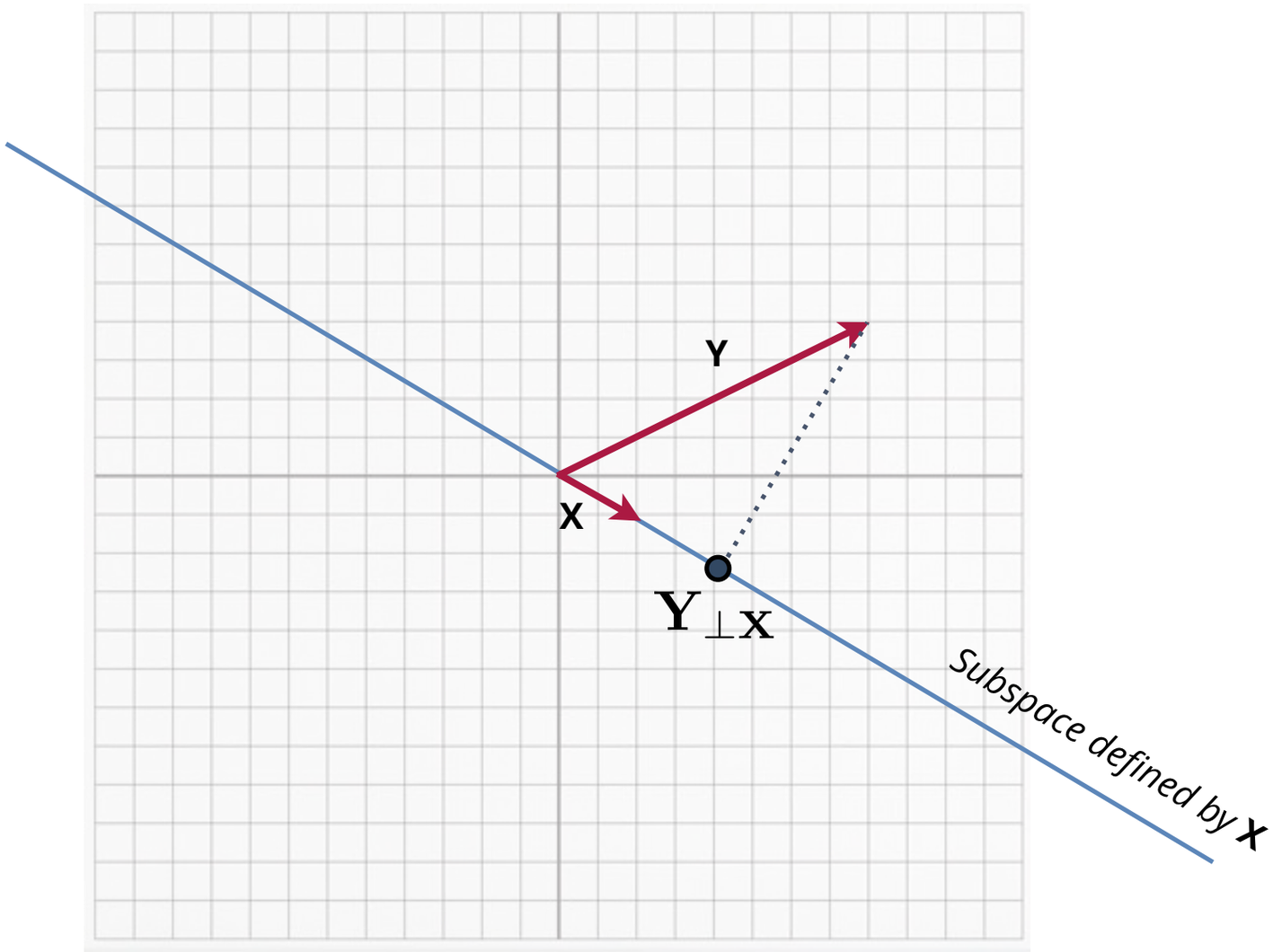
Odysseus leaves from Aeolus Island and wants to go home to Ithaca. The winds, Zephyros blows his ship off course. Where should Odysseus jump ship and swim to shore? The shortest swim is when his ship is orthogonal to Ithaca. This is the projection of Ithaca onto the subspace defined by where Zephyros blew his ship.

This metaphor for the idea of projection comes from Daniel Kaplan.

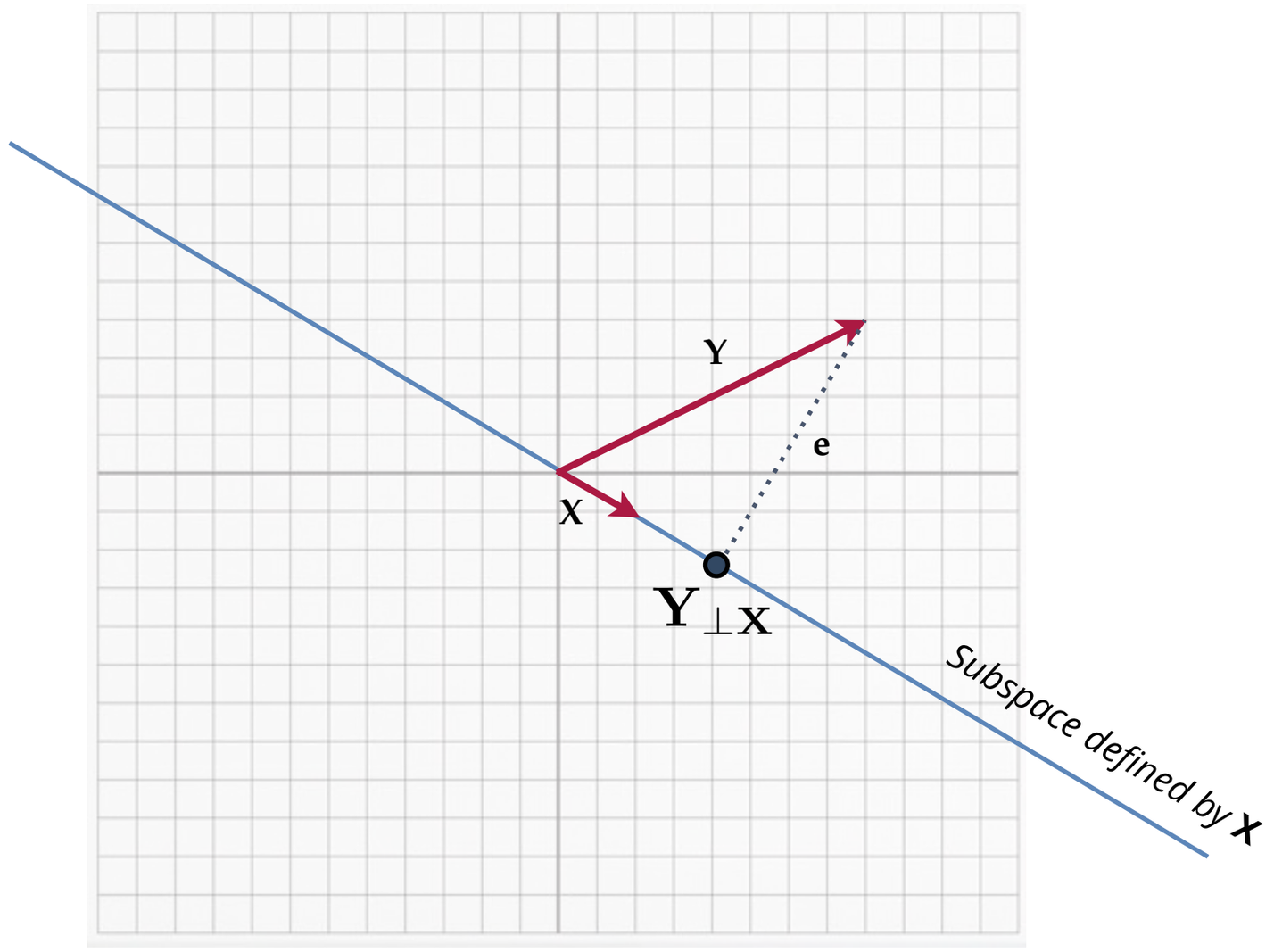
Statistical Modeling and Projection

Fitting the model Y ~ 1 + X is akin to the finding the projection of **Y** onto the subspace of **X**, where **X** is the design matrix.

The subspace of **X** typically has more than two dimensions but is shown as such for easy visualization.



This is similar to our earlier metaphor. A model's goal is to use the data in **X** to reproduce the values in **Y**. Linear combinations of **X** can produce any value on the subspace. The closest we get to **Y** is the point produced by a linear combination of **X** that is at $\mathbf{Y}_{\perp \mathbf{X}}$. The distance to **Y** from this point is residual.



Response Variable

Subspace of the Explanatory Variables

Residuals

The projection of **Y** on to the subspace of **X** is the linear combination expressed by the fitted model,

$$\mathbf{Y}_{\perp \mathbf{X}} = \beta \mathbf{X}$$

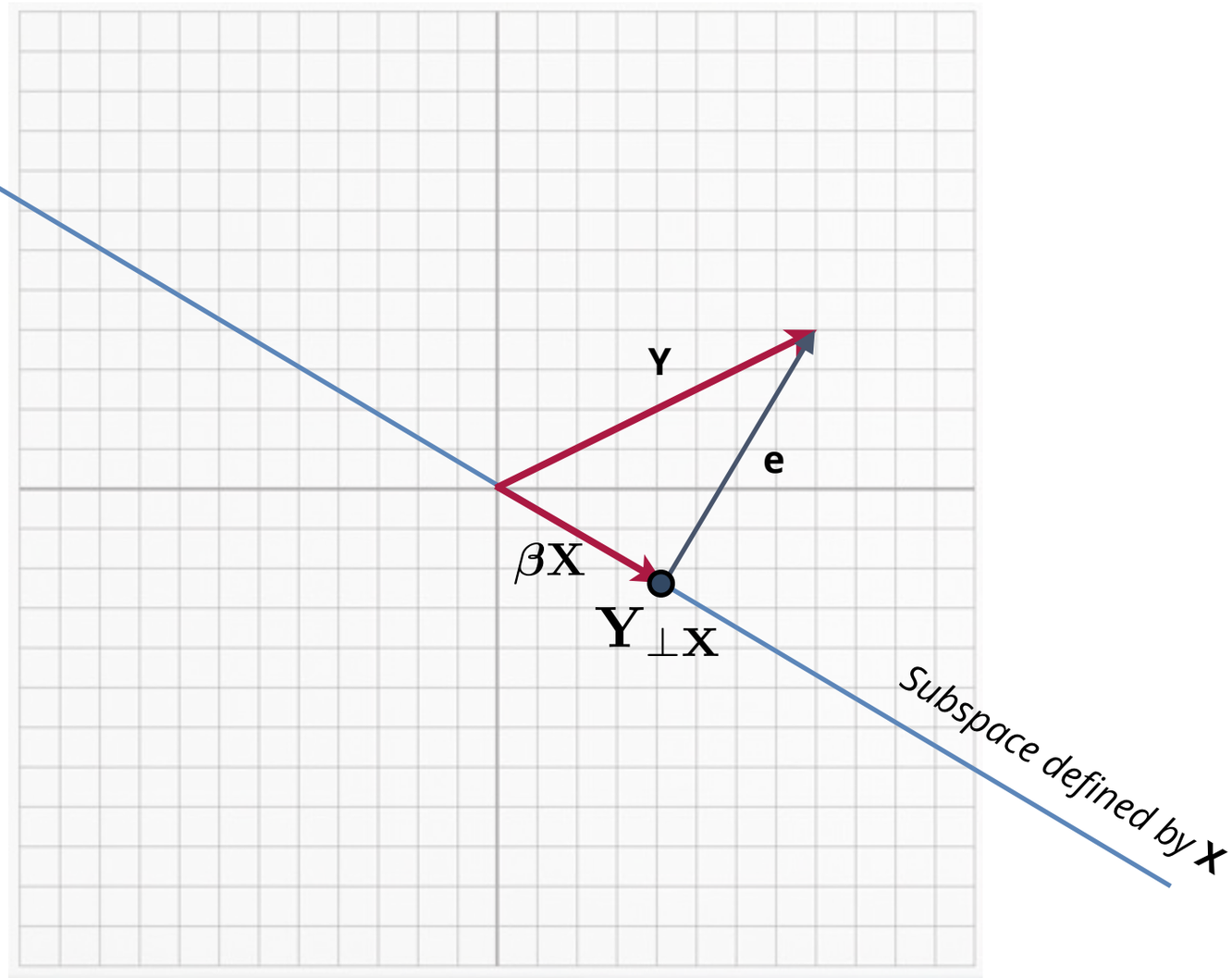
The residual is the vector between the point $\mathbf{Y}_{\perp \mathbf{X}}$ and the goal vector **Y**.

Using ideas of vector geometry,

$$\mathbf{Y} = \beta \mathbf{X} + \mathbf{e}$$

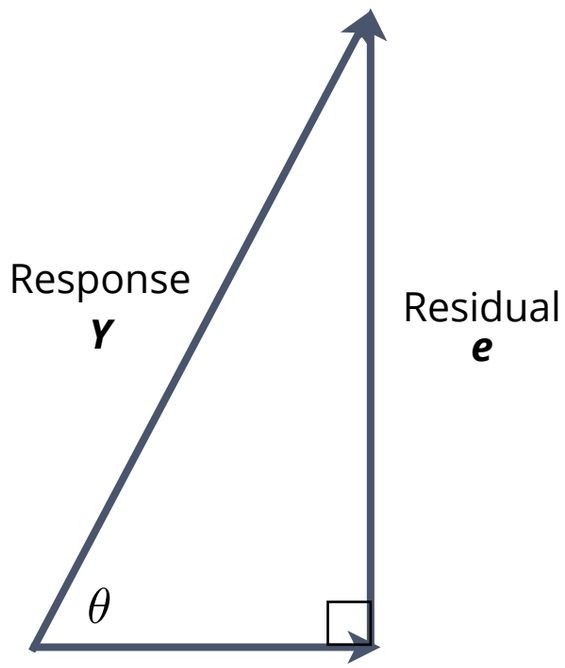
Computing the residual vector is simply vector subtraction

$$\mathbf{e} = \mathbf{Y} - \beta \mathbf{X}$$



Model Triangle

Model Triangle

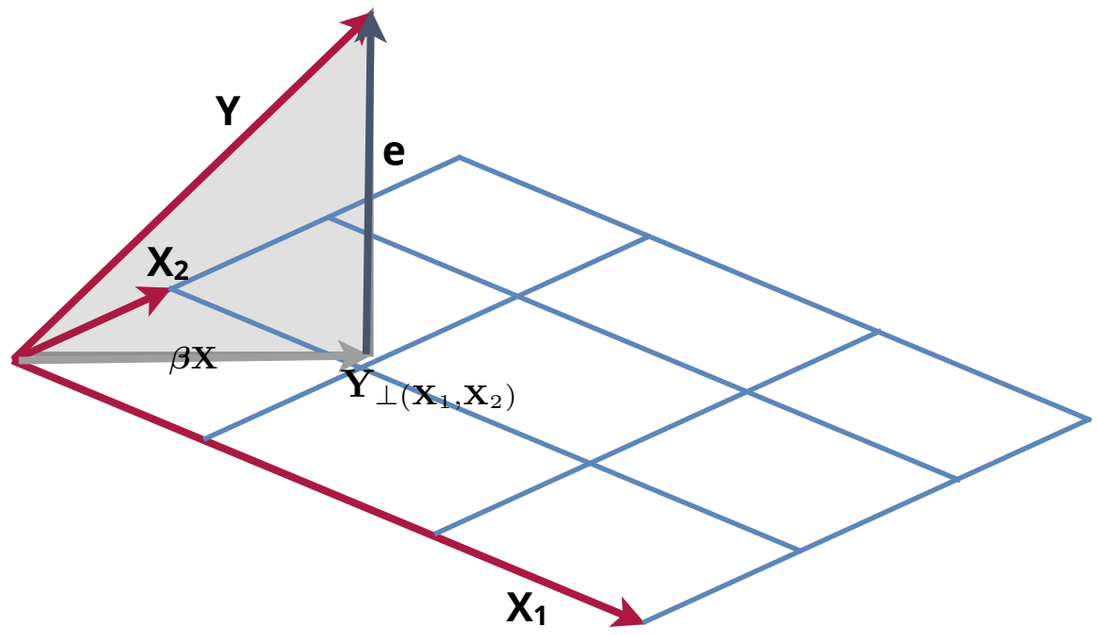


Response = Fitted values + Residuals

These vectors will always form a triangle because of the vector arithmetic that computes the residuals as the difference between the response vector and the fitted model vector. Furthermore, the residual vector is always **perpendicular** (orthogonal) to the fitted model vector.

Fitted Values $\beta\mathbf{X}$ $\quad \mathbf{Y}_{\perp\mathbf{X}}$

In three dimensions, the model triangle is formed by the projection onto the subspace of the **plane** consisting of all linear combinations of X_1 and X_2. But the relationships still hold.



Example of the Model Triangle: Mean of Y

Finding the mean is equivalent to fitting the intercept-only model, Y ~ 1

Let $\mathbf{Y} = \begin{bmatrix} 5 \\ 1 \end{bmatrix}$ then, $\bar{y} = 3,\ \hat{\sigma}_y = 2.83$

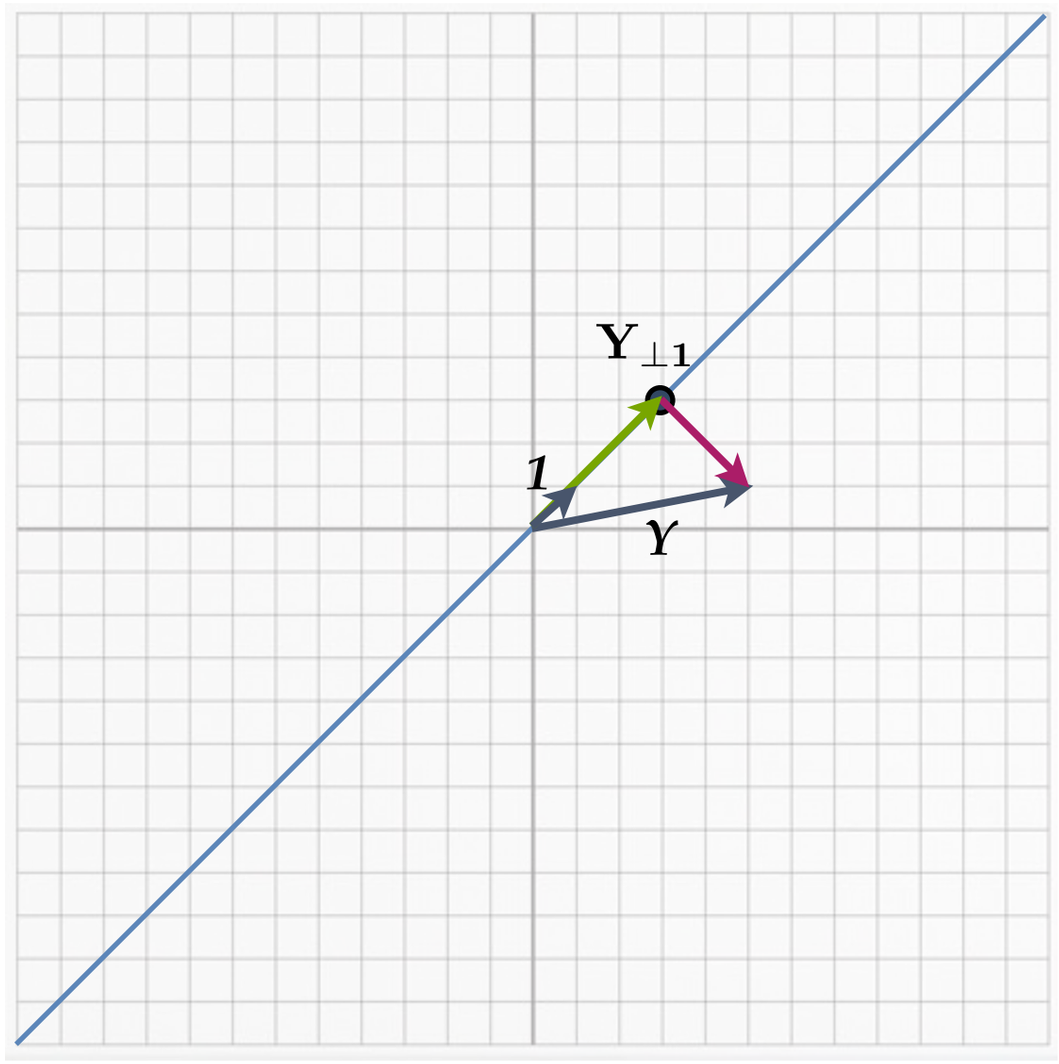
The intercept vector is a vector of ones.

$$\mathbf{1} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Find the coefficient, ***c***, such that $\mathbf{Y}_{\perp \mathbf{1}} = c\mathbf{X}$

$$\mathbf{Y}_{\perp \mathbf{1}} = 3 \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$$

The mean corresponds to the fitted model vector.



Sum of Squares

Since the model triangle is a right triangle, the Pythagorean Theorem relates the lengths of the three sides as:

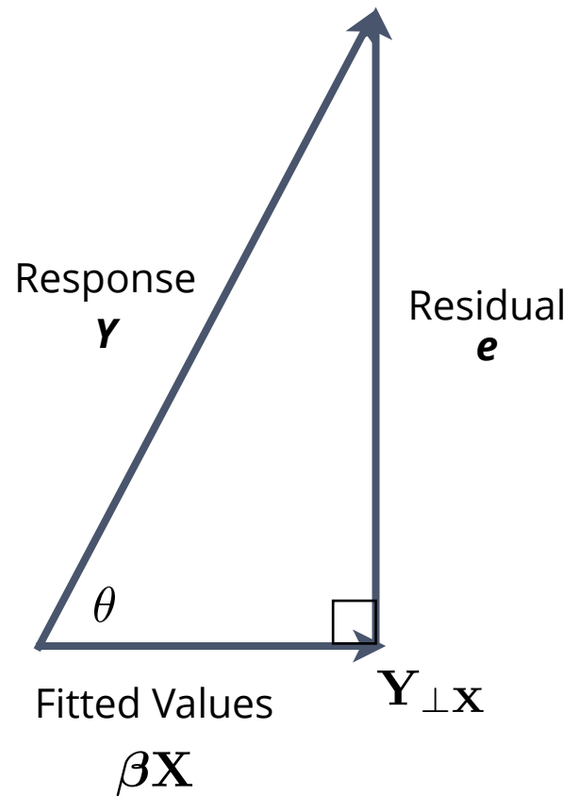
$$\|\mathbf{Y}\|^2 = \|\beta\mathbf{X}\|^2 + \|\mathbf{e}\|^2$$

This can be expressed as dot products:

$$\mathbf{Y} \bullet \mathbf{Y} = \beta\mathbf{X} \bullet \beta\mathbf{X} + \mathbf{e} \bullet \mathbf{e}$$

A vector dotted with itself is just the sum of each element squared (a sum of squares):

$$\mathrm{SS}_{\mathrm{Y}} = \mathrm{SS}_{\mathrm{Model}} + \mathrm{SS}_{\mathrm{Residual}}$$



Mean and SD via Vectors

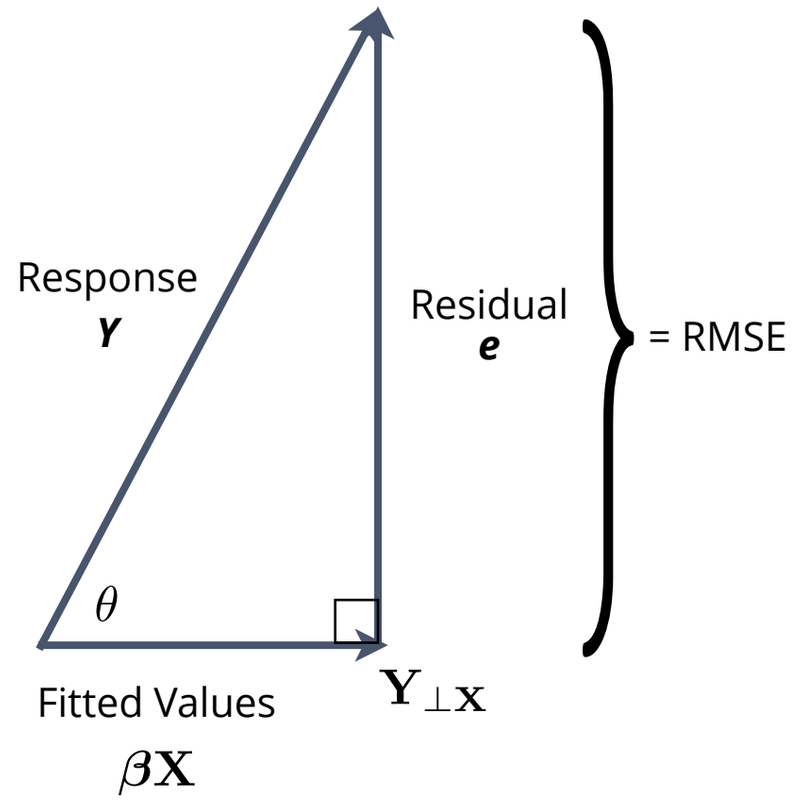
RMSE/Standard Deviation (Example Cntd.)

The residual vector can be computed by subtracting the fitted vector from the response vector.

$$\mathbf{Y} - \boldsymbol{\beta}\mathbf{X} = \begin{bmatrix} 5 \\ 1 \end{bmatrix} - \begin{bmatrix} 3 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ -2 \end{bmatrix}$$

The **length of this residual vector** is the Root Mean Square Error (RMSE).

$$\begin{aligned} ||\mathbf{e}|| &= \sqrt{\mathbf{e} \bullet \mathbf{e}} \\ &= \sqrt{\begin{bmatrix} 2 \\ -2 \end{bmatrix} \bullet \begin{bmatrix} 2 \\ -2 \end{bmatrix}} \\ &= \sqrt{8} \\ &= 2.83 \end{aligned}$$



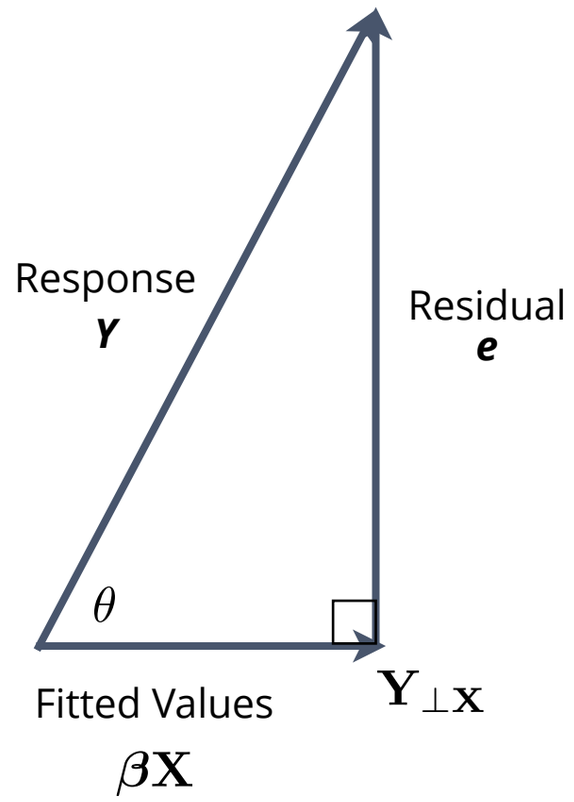
In the intercept-only model, the RMSE is equivalent to the standard deviation of *Y*.

Variation Accounted For/Multiple Correlation

Since the model triangle is a right triangle, we can use trigonometry to also relate the side lengths

$$\cos(\theta) = \frac{\text{Adjacent}}{\text{Hypotenuse}}$$

$$\cos(\theta) = \frac{\|\beta\mathbf{X}\|}{\|\mathbf{Y}\|} = \frac{\sqrt{\beta\mathbf{X} \bullet \beta\mathbf{X}}}{\sqrt{\mathbf{Y} \bullet \mathbf{Y}}}$$



Squaring both sides of the equation...

$$\cos(\theta)^2 = \frac{\beta\mathbf{X} \bullet \beta\mathbf{X}}{\mathbf{Y} \bullet \mathbf{Y}}$$

The numerator and denominator are both sum of squares!

$$\cos(\theta)^2 = \frac{\text{SS}_{\text{Model}}}{\text{SS}_{\text{Total}}} = R^2 \quad \text{Which implies...} \quad \cos(\theta) = R$$

The **cosine of the angle between the fitted vector and the response vector** is the correlation between the fitted values and the response, which with only one predictor, is the correlation between *X* and *Y*.