Vector Geometry for Statistical Models

Statistical Geometry

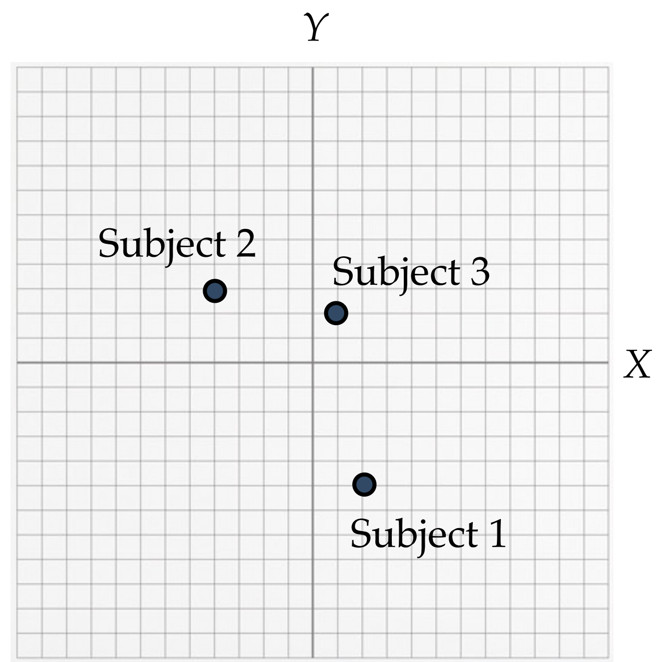
It is possible to compute all statistical calculations for regression models using only a ruler and protractor.

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

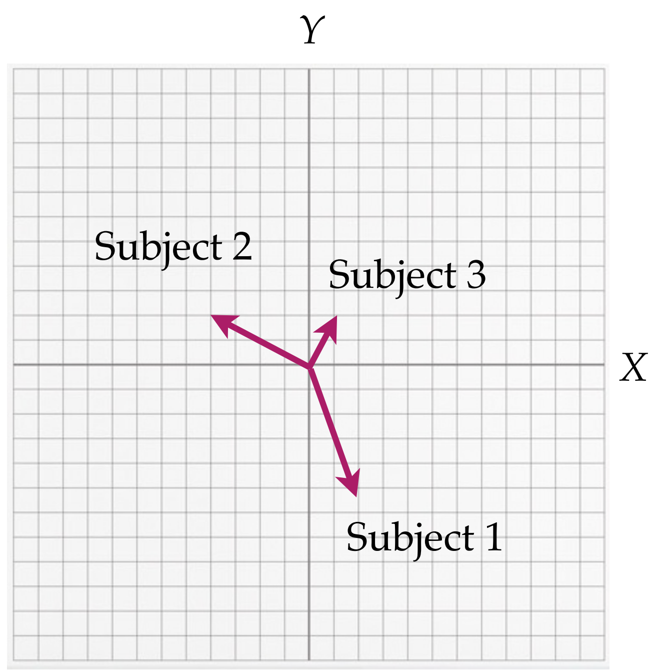
We typically plot multivariate data in variable space. In variable space, the **variables are represented as axes** and the **subjects are represented as points**, which are plotted based on their values for the variables.

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.

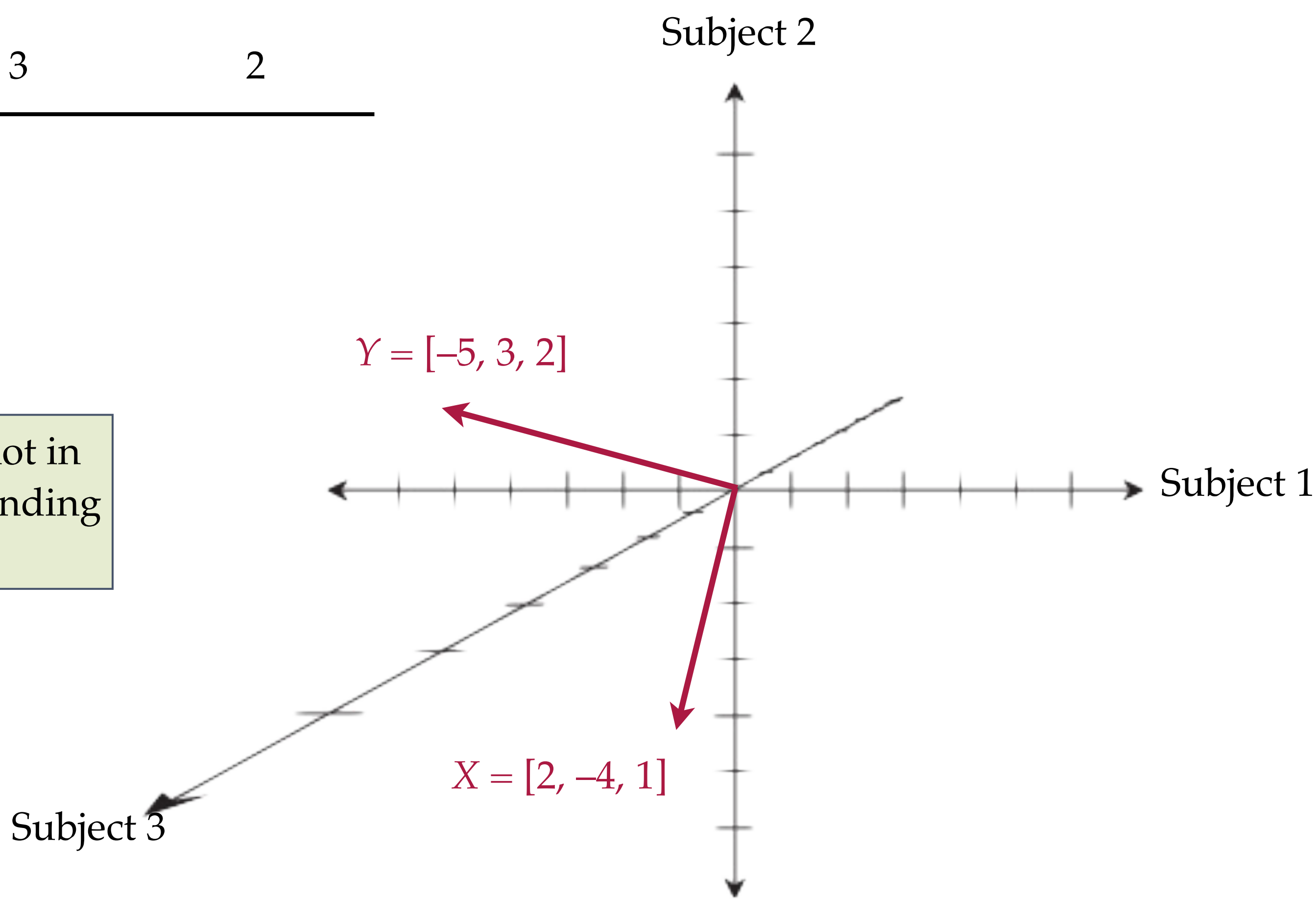
Subject	**X**	**Y**
Subject 1	2	–5
Subject 2	–4	3
Subject 3	1	2



Subject Space

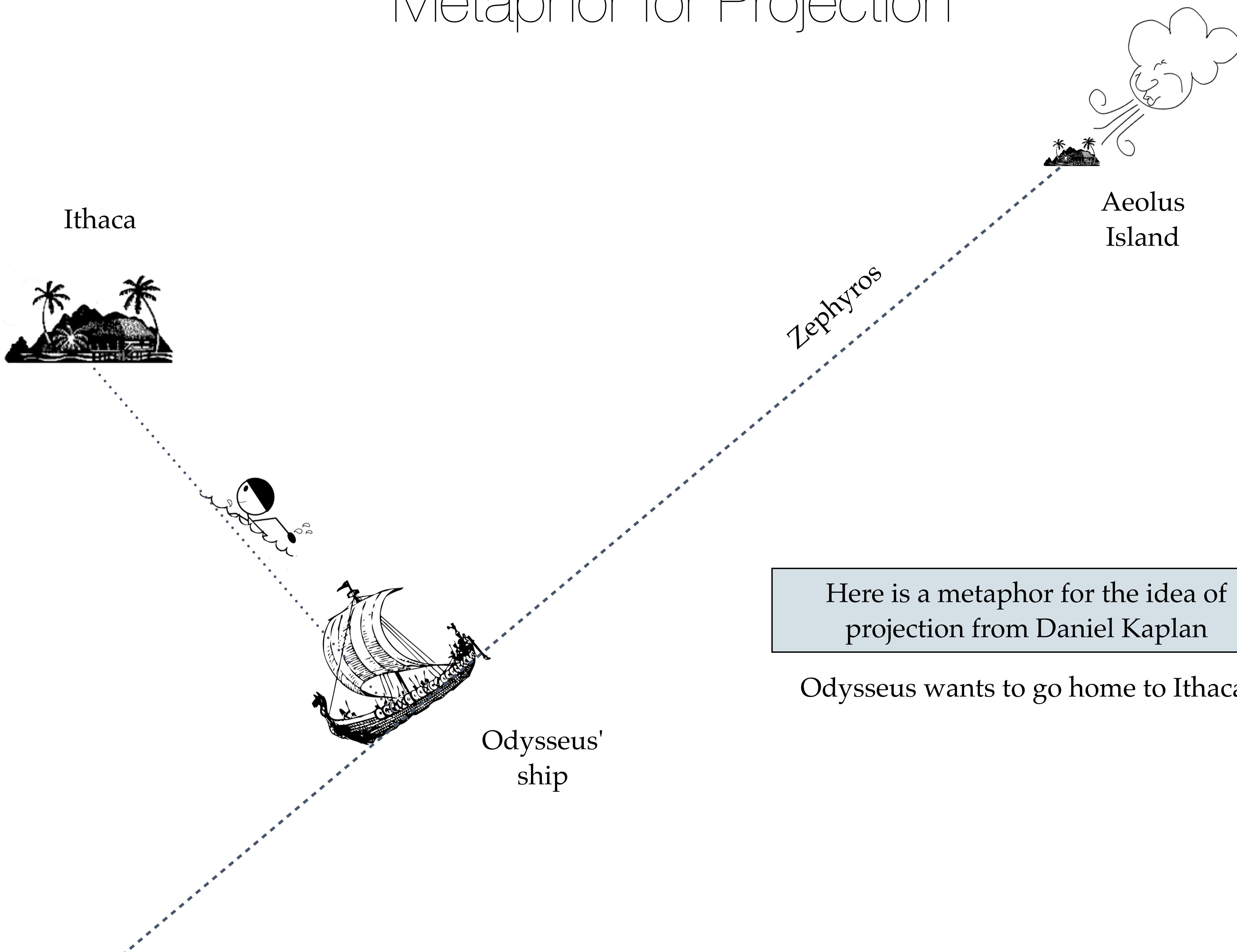
Variable	Subject 1	Subject 2	Subject 3
X	2	–4	1
Y	–5	3	2

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.

Metaphor for Projection

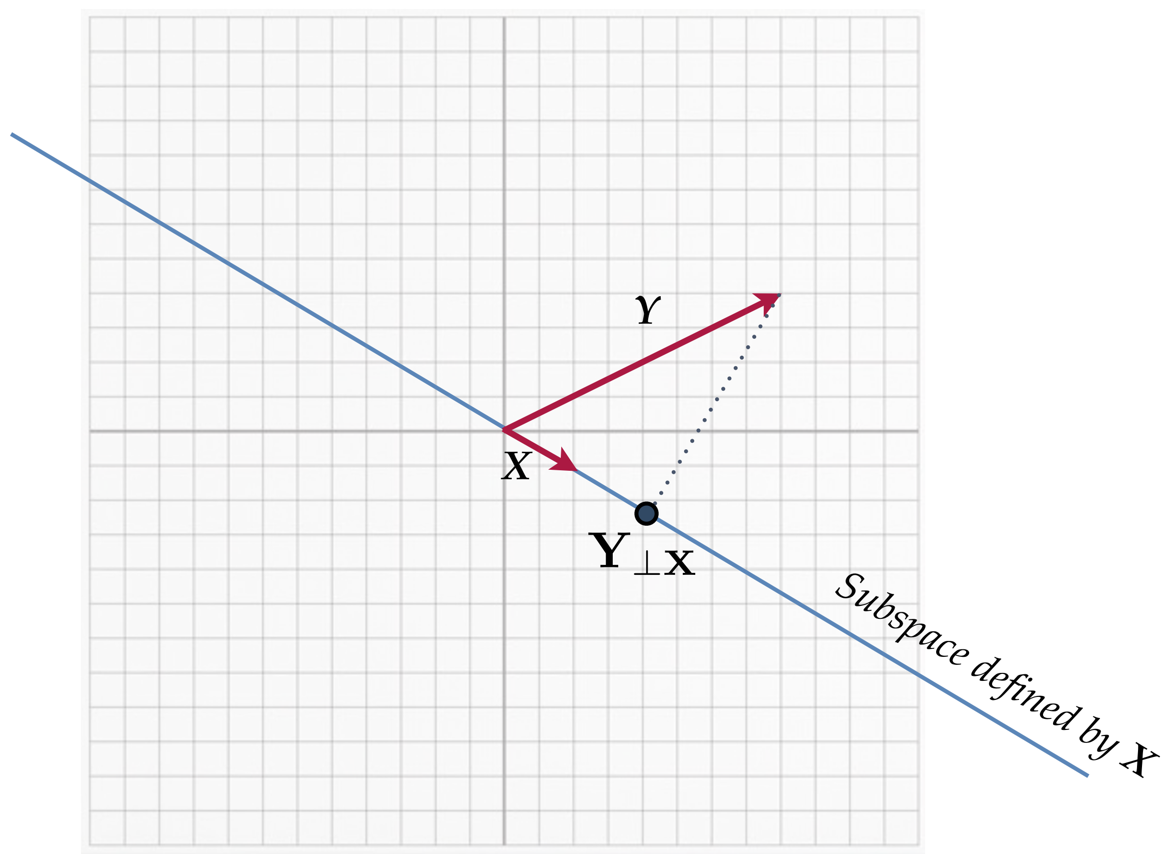


Here is a metaphor for the idea of projection from Daniel Kaplan

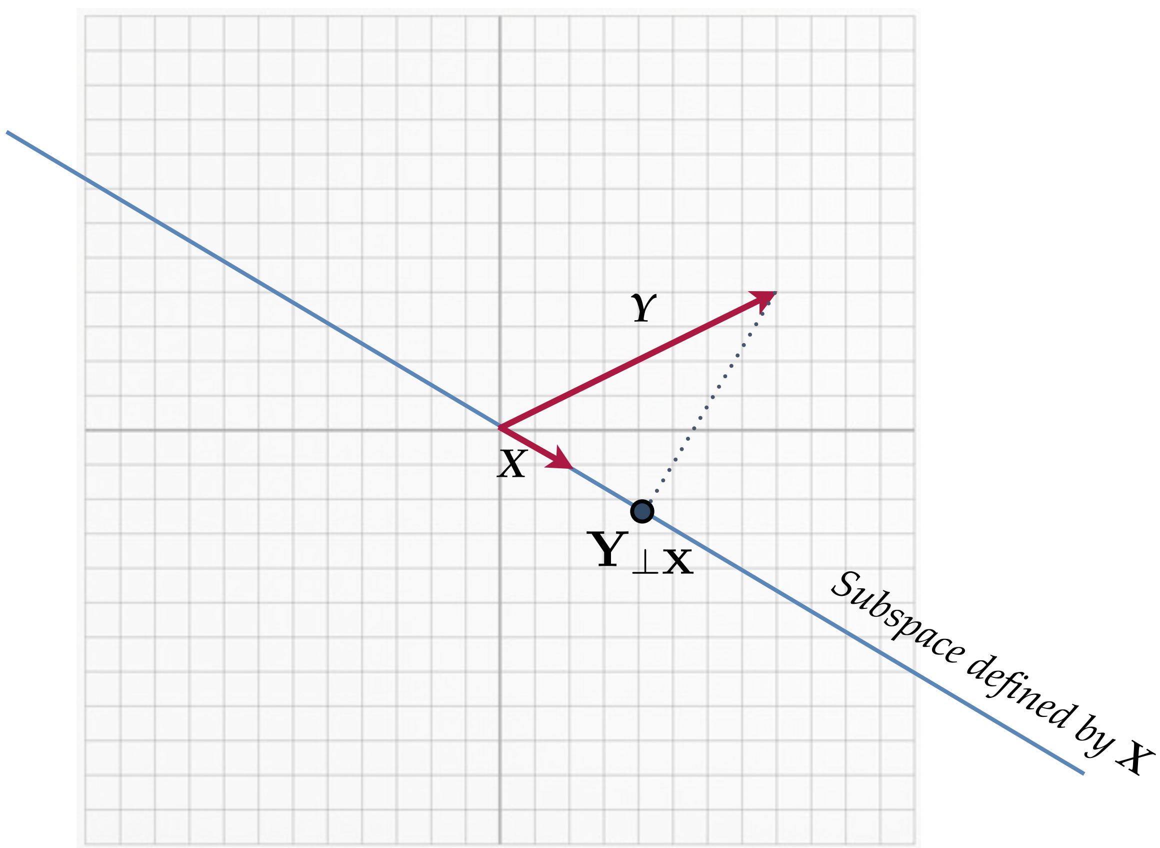
Odysseus wants to go home to Ithaca.

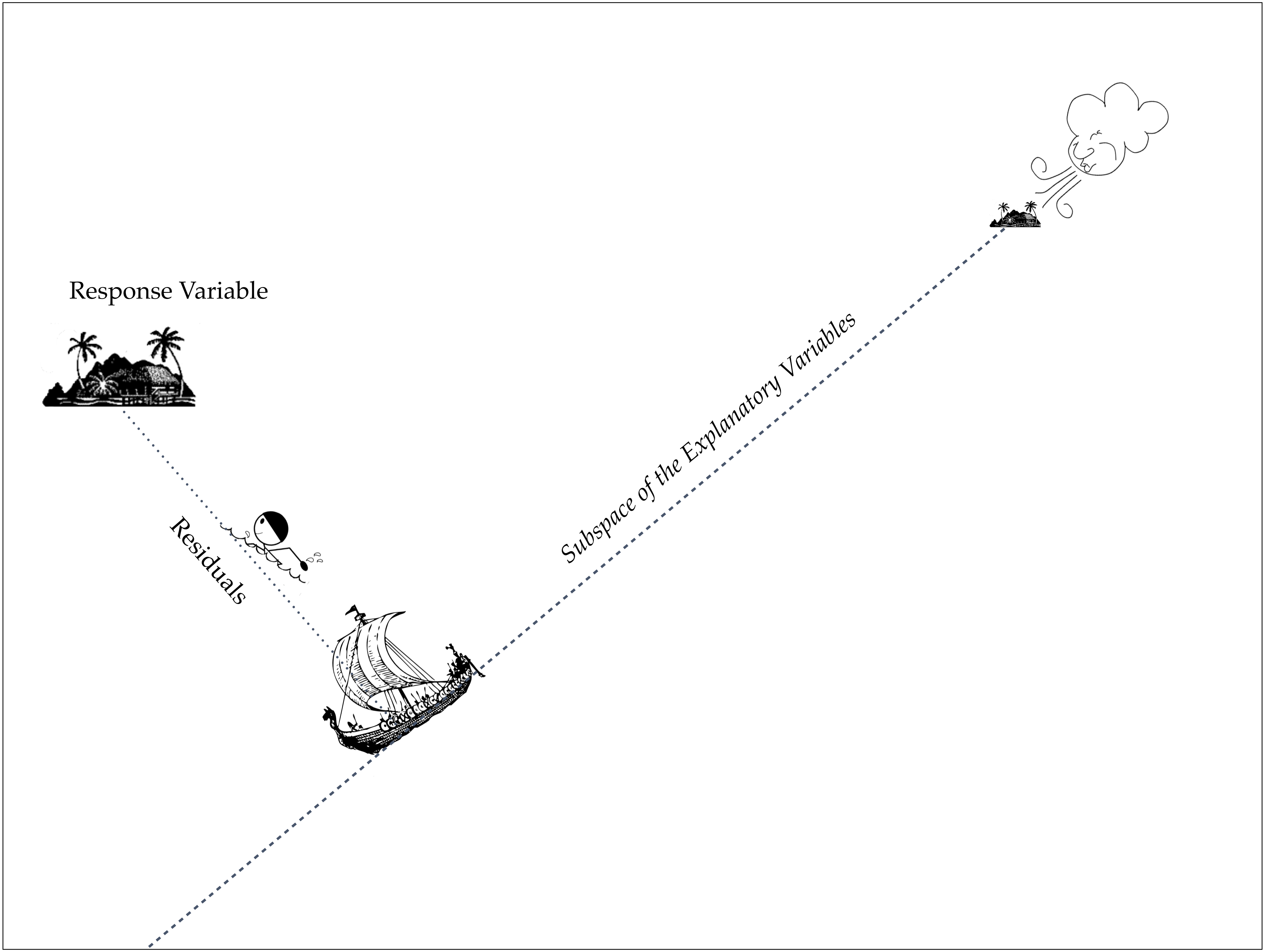
Statistical Modeling and Projection

Fitting the model `Y ~ X` is akin to the finding the projection of **Y** onto the subspace of **X**



This is similar to our earlier metaphor. We start at the origin and our goal is to get to $\boldsymbol{Y}$. The winds blows us in the direction of $\boldsymbol{X}$ along the subspace. The closest we get to $\boldsymbol{Y}$ is the point $\mathbf{Y}_{\perp\mathbf{X}}$. The remaining part of the journey (the swimming) is the residual part of the journey.

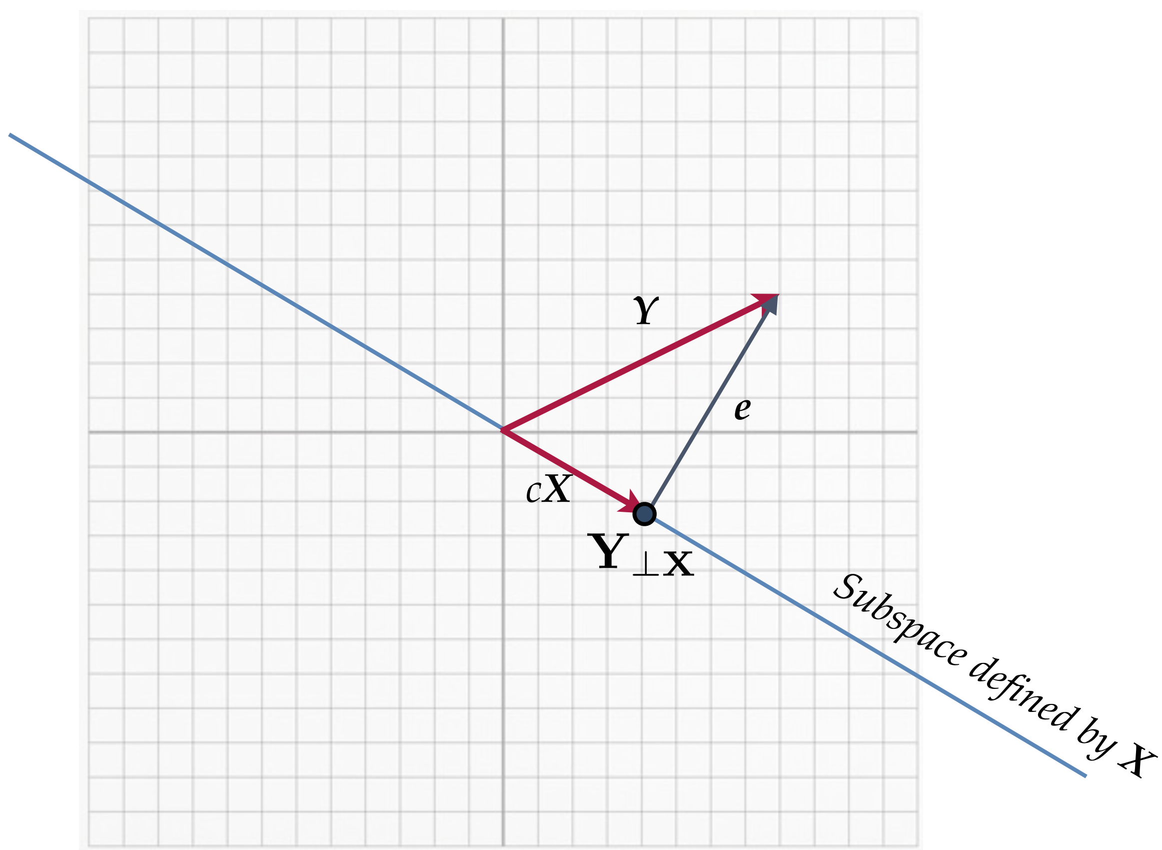



Response Variable
Residuals
Subspace of the Explanatory Variables

Once the projected point $\mathbf{Y}_{\perp\mathbf{X}}$ has been found, this can be translated to a coefficient (the scalar that extends X to this point).

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

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



Since

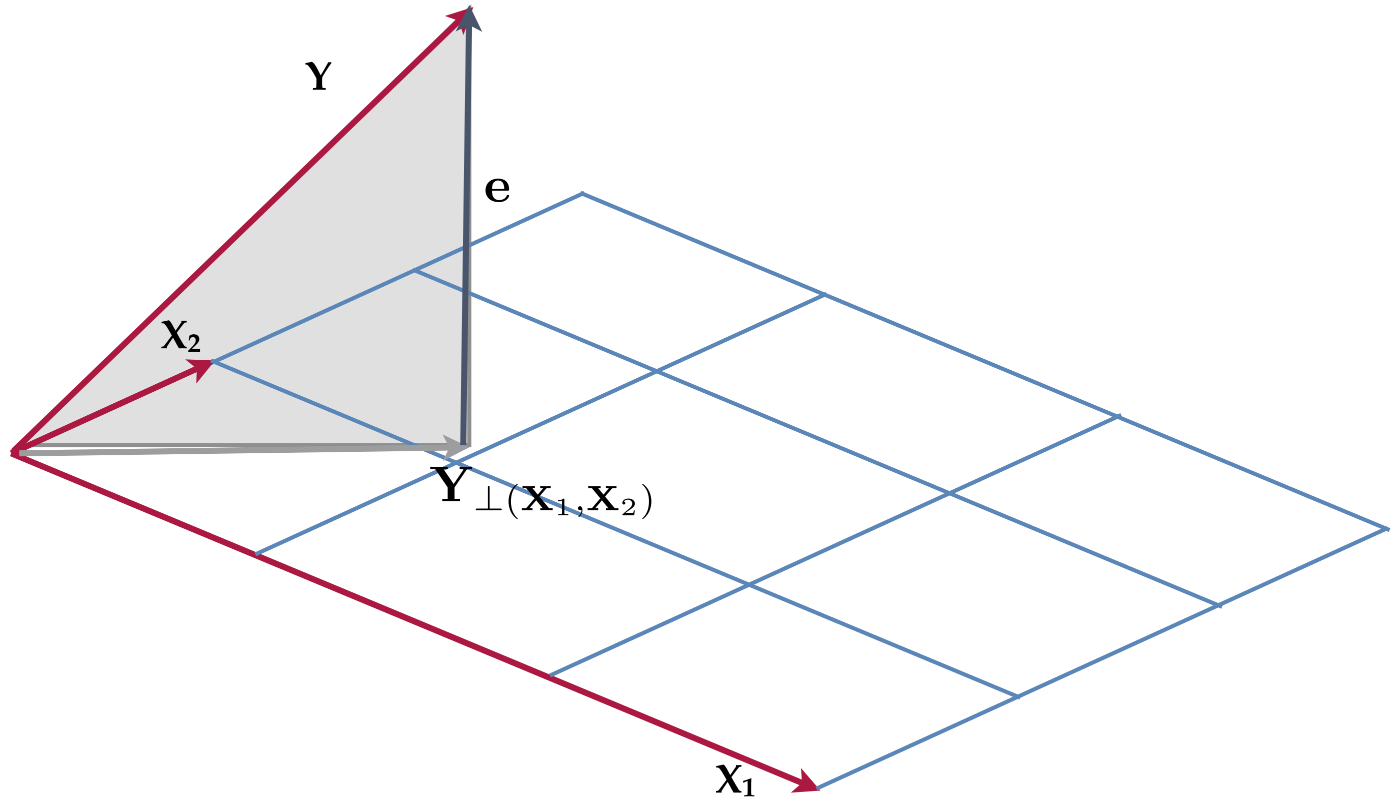
$$\mathbf{Y}_{\perp\mathbf{X}} + \mathbf{e} = \mathbf{Y}$$

Computing the residual vector is simply vector subtraction

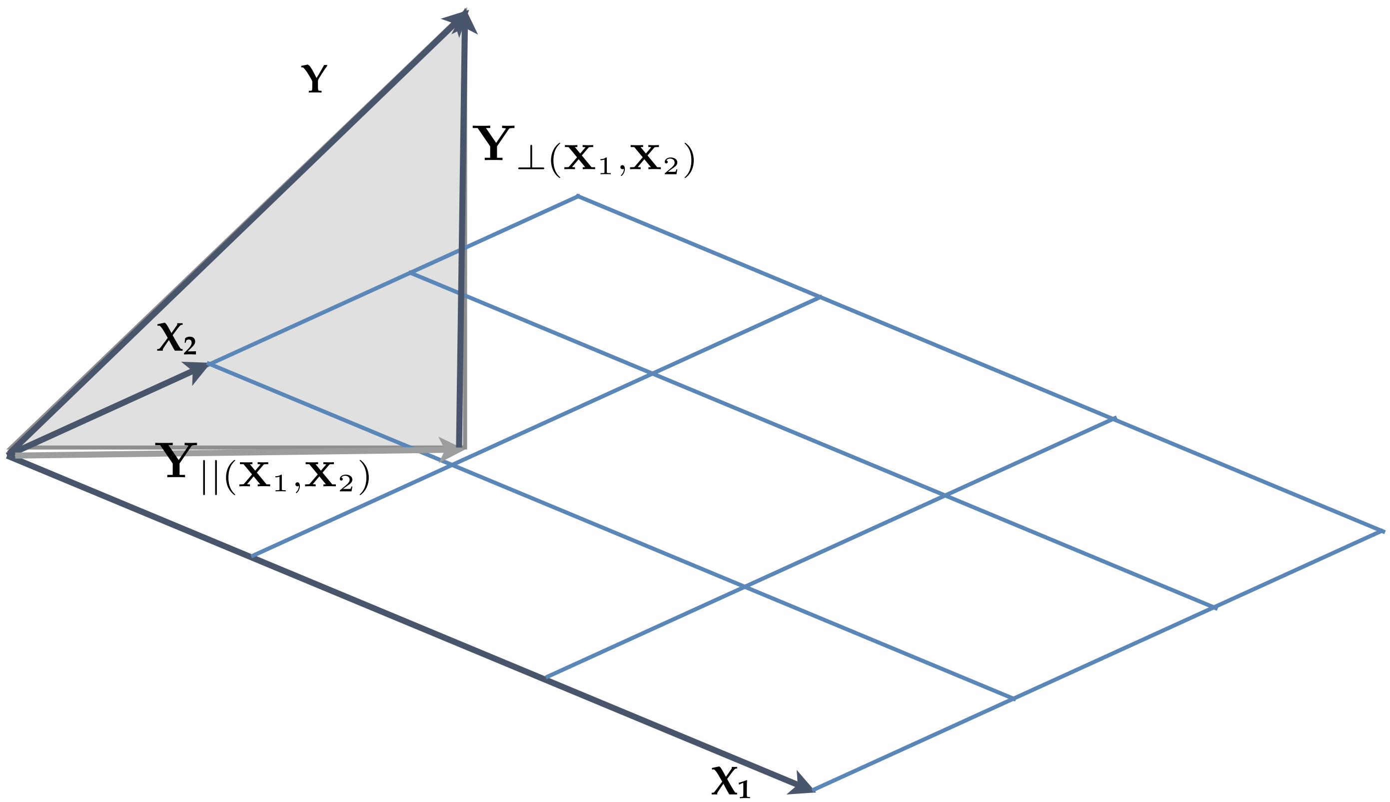
$$\mathbf{e} = \mathbf{Y} - \mathbf{Y}_{\perp\mathbf{X}}$$

Multiple Predictors

Consider the model `Y ~ X1 + X2`. The subspace of two explanatory variables is a **plane** consisting of all linear combinations of X_1 and X_2.



The coefficients for $\mathbf{X}_1$ and $\mathbf{X}_2$ give the linear combination that scale each explanatory variable to reach the projected point from $\mathbf{Y}$ on the ($\mathbf{X}_1$, $\mathbf{X}_2$)-subspace.

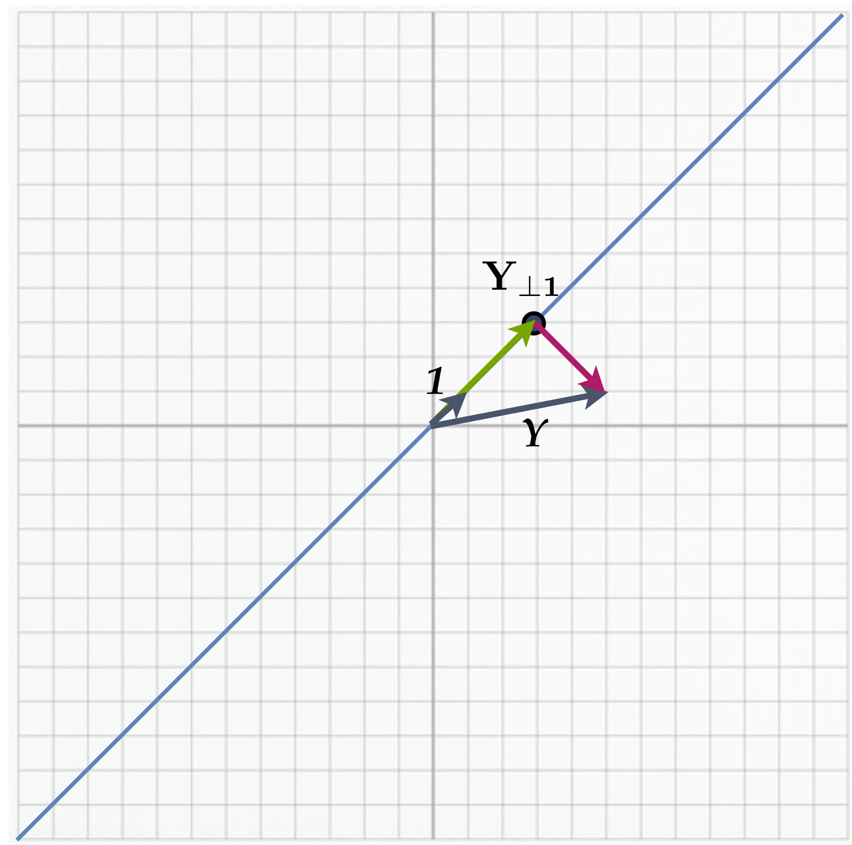


Mean and SD via Vectors

Finding the Mean

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.

Residuals

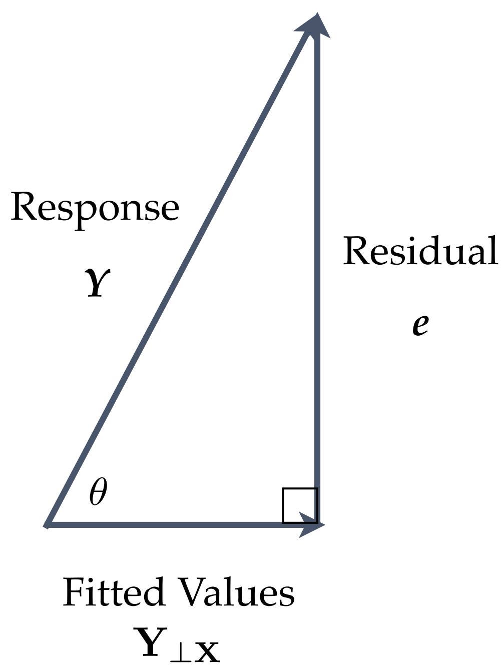
$$\mathbf{Y} - \mathbf{Y}_{\perp\mathbf{1}} = \begin{bmatrix} 5 \\ 1 \end{bmatrix} - \begin{bmatrix} 3 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ -2 \end{bmatrix}$$

The **length of this residual vector** is the standard deviation of Y.

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

No good reason to use geometry to compute mean or standard deviation, but it does emphasize the fact that the mean and standard deviation are two complimentary aspects of a variable.

Model Triangle



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** to the fitted model vector.

Response = Fitted values + Residuals

Sum of Squares

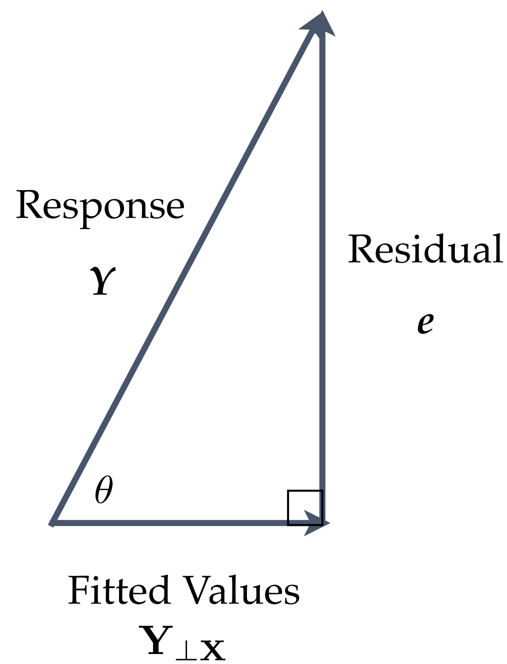
Since the model triangle is a right triangle, the Pythagorean Theorem relates the lengths of the three sides as

$$||\mathbf{Y}||^2 = ||\mathbf{Y}_{\perp \mathbf{x}}||^2 + ||\mathbf{e}||^2$$

$$\mathbf{Y} \bullet \mathbf{Y} = \mathbf{Y}_{\perp \mathbf{x}} \bullet \mathbf{Y}_{\perp \mathbf{x}} + \mathbf{e} \bullet \mathbf{e}$$

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

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



Correlation

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

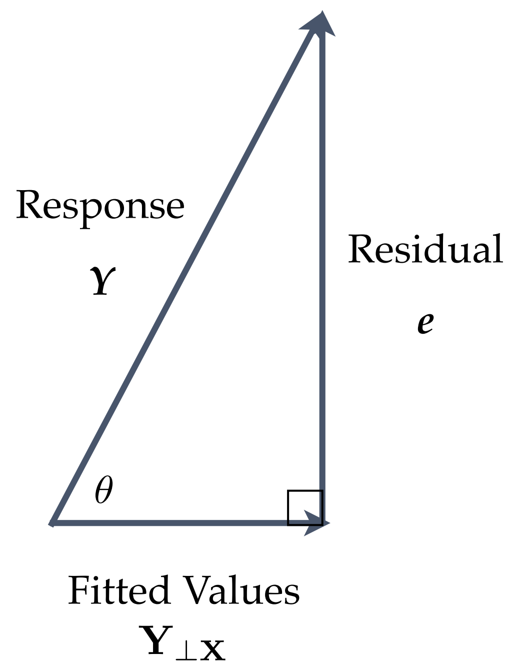
$$\cos\theta = \frac{\text{Adjacent}}{\text{Hypotenuse}}$$

$$\cos\theta = \frac{||\mathbf{Y}_{\perp\mathbf{x}}||}{||\mathbf{Y}||} = \frac{\sqrt{\mathbf{Y}_{\perp\mathbf{x}} \bullet \mathbf{Y}_{\perp\mathbf{x}}}}{\sqrt{\mathbf{Y} \bullet \mathbf{Y}}}$$

Squaring both sides of the equation...

$$\big[\cos\theta\big]^2 = \frac{\mathbf{Y}_{\perp\mathbf{x}} \bullet \mathbf{Y}_{\perp\mathbf{x}}}{\mathbf{Y} \bullet \mathbf{Y}}$$

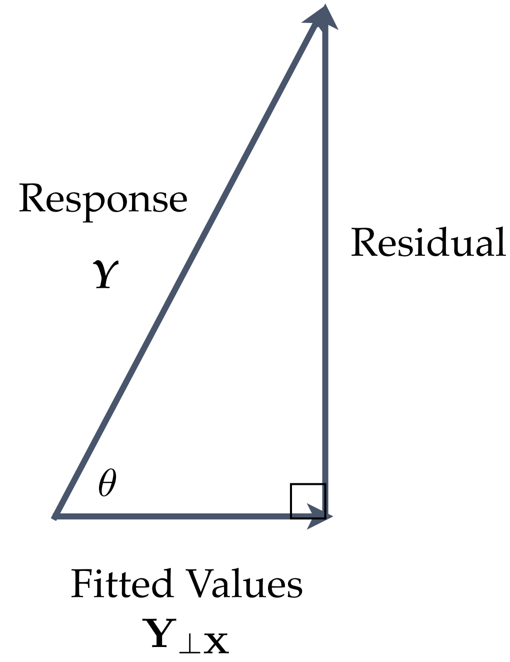
The numerator and denominator are both sum of squares!



$$\left[\cos\theta\right]^2 = \frac{\mathrm{SS}_{\mathrm{Model}}}{\mathrm{SS}_{\mathrm{Total}}}$$

$$\left[\cos\theta\right]^2 = R^2$$

$$\cos\theta = r$$



The **cosine of the angle between the fitted model 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.