Vectors and Statistical Models

Data

Consider using this "toy" dataset to model the variation in wages

```
> toy = read.csv(file = "toyData.csv")
> toy

  wage educ sex  status age   sector
 12.00   12   M Married  32    manuf
  8.00   12   F Married  33  service
 16.26   12   M  Single  32  service
 13.65   16   M Married  33     prof
  8.50   17   M  Single  26 clerical
```

Both the outcome (wage) and some of the predictors are quantitative (educ, age), and others are categorical (sex, status, sector)

In building the model you could include main-effects and/or interactions between predictors

Model Vectors: Quantitative Variables

Model vectors are the translation of the model terms into vectors. They are also referred to as **indicator variables**.

Each **quantitative variables** has a *single* model vector or indicator. For the three quantitative variables in our example, the model vectors are:

wage

$$\begin{bmatrix} 12.00 \\ 8.00 \\ 16.26 \\ 13.65 \\ 8.50 \end{bmatrix}$$

educ

$$\begin{bmatrix} 12 \\ 12 \\ 12 \\ 16 \\ 17 \end{bmatrix}$$

age

$$\begin{bmatrix} 32 \\ 33 \\ 32 \\ 33 \\ 26 \end{bmatrix}$$

Model Vectors: Categorical Variables

Categorical variables (factors) have *multiple* model vectors (indicators). There is **one model vector per level** of the factor.

The sex main-effect would be composed of two model vectors.

$$\texttt{sexF}: \begin{bmatrix}0\\1\\0\\0\\0\end{bmatrix} \quad \texttt{sexM}: \begin{bmatrix}1\\0\\1\\1\\1\end{bmatrix}$$

The status main-effect would also be composed of two model vectors.

$$\texttt{statusMarried}: \begin{bmatrix}1\\1\\0\\1\\0\end{bmatrix} \quad \texttt{statusSingle}: \begin{bmatrix}0\\0\\1\\0\\1\end{bmatrix}$$

The sector main-effect would be composed of five model vectors.

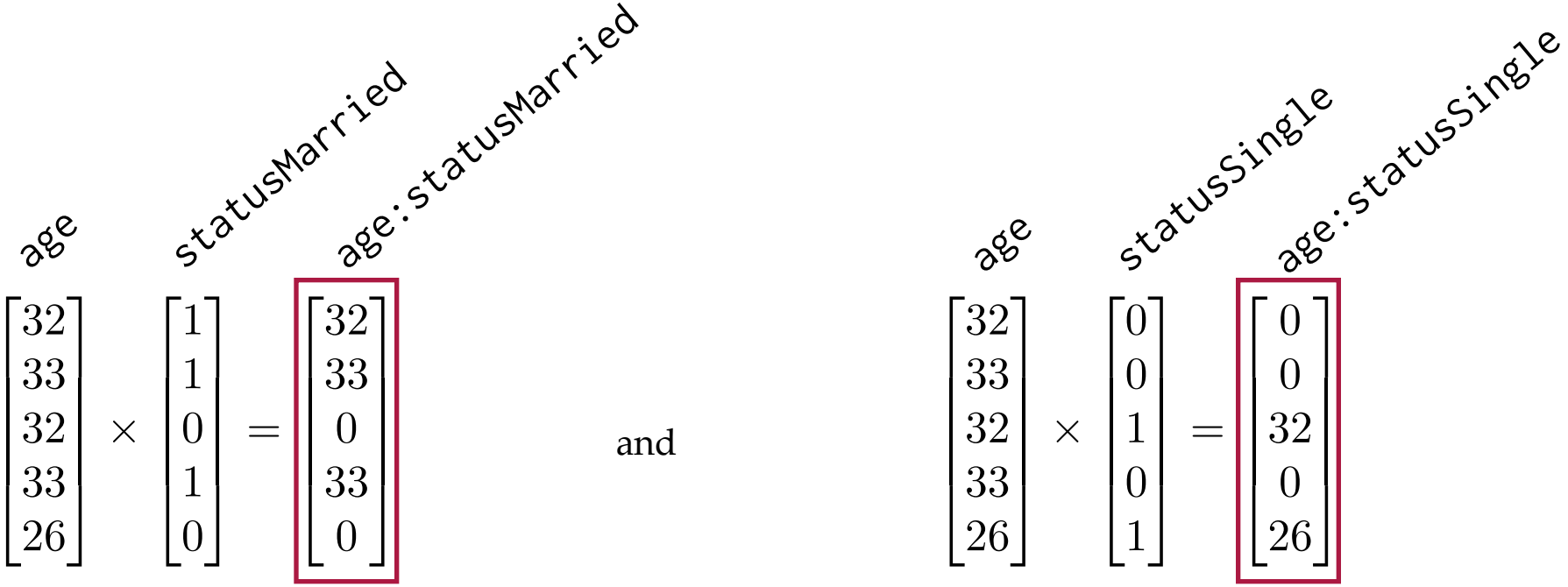
$$\texttt{sectorclerical}: \begin{bmatrix}0\\0\\0\\0\\1\end{bmatrix} \quad \texttt{sectorconstr}: \begin{bmatrix}0\\0\\0\\0\\0\end{bmatrix} \quad \texttt{sectormanuf}: \begin{bmatrix}1\\0\\0\\0\\0\end{bmatrix} \quad \texttt{sectorprof}: \begin{bmatrix}0\\0\\0\\1\\0\end{bmatrix} \quad \texttt{sectorservice}: \begin{bmatrix}0\\1\\1\\0\\0\end{bmatrix}$$

Model Vectors: Interaction Terms

The model vectors for any interaction terms included are the pairwise products between all of the model vectors for the predictors included in the interaction.

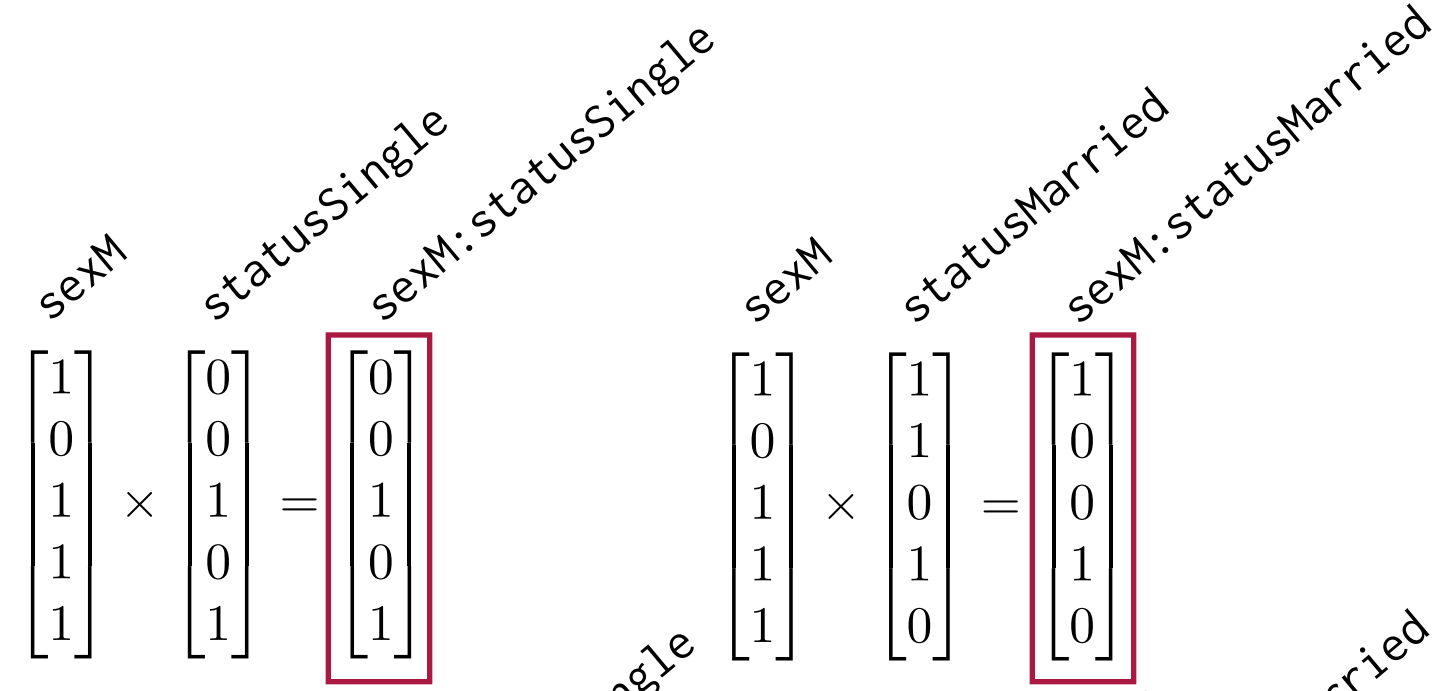
$$\underset{\text{educ}}{\begin{bmatrix}12\\12\\12\\16\\17\end{bmatrix}} \times \underset{\text{age}}{\begin{bmatrix}32\\33\\32\\33\\26\end{bmatrix}} = \underset{\text{educ:age}}{\begin{bmatrix}384\\396\\384\\528\\422\end{bmatrix}}$$

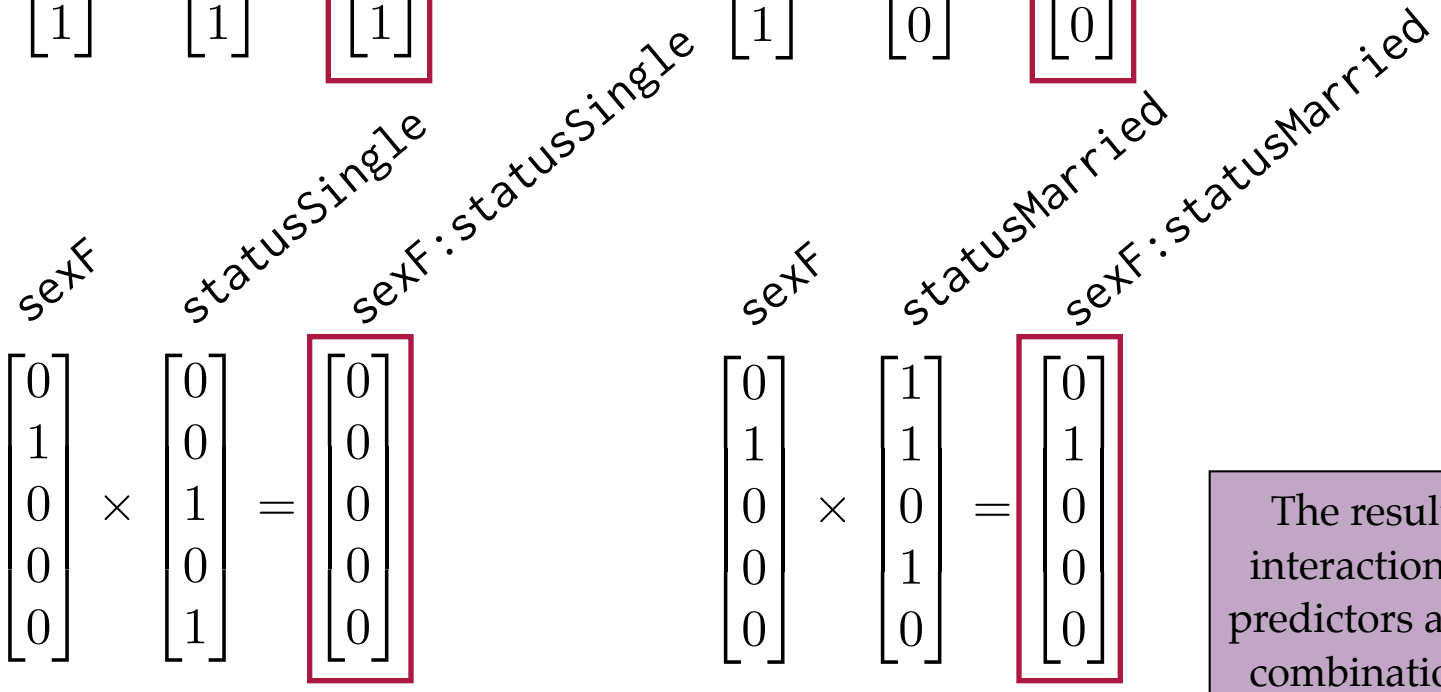
There is only one model vector for the interaction between two quantitative predictors.



There would be two model vectors for the interaction between age and marital status.

There would be four model vectors for the interaction between sex and marital status.





The resulting model vectors for the interaction terms between categorical predictors are indicators of the different combinations of the initial predictors.

Model Vectors: Intercept

The intercept is a special model vector of all ones.

$$\texttt{Intercept}$$

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

Model Matrix

When constructing a model, you supply a list of model terms, such as

```
lm.1 = lm(wage ~ 1 + age + sex, data = toy)
```

The software (e.g., R) then translates each of the predictors to their model vectors

$$\begin{matrix} \texttt{Intercept} & \texttt{age} & \texttt{sexM} & \texttt{sexF} \end{matrix}$$
$$\begin{bmatrix} 1 & 32 & 1 & 0 \\ 1 & 33 & 0 & 1 \\ 1 & 32 & 1 & 0 \\ 1 & 33 & 1 & 0 \\ 1 & 26 & 1 & 0 \end{bmatrix}$$

A set of vectors having the same dimension is called a *matrix*. The matrix composed of all the model vectors is called the **model matrix** or **design matrix**.

When an intercept is included in the model, not all of the indicators for categorical predictors are used!

Fitted Values

```
> summary(lm.1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.5402    13.6620  -1.430    0.289
age           0.8346     0.4083   2.044    0.178
sexM          6.4802     2.6928   2.407    0.138
```

To obtain fitted values, the model vectors are scaled by the coefficients (scalars) and added. The fitted values are a **linear combination** of the model vectors.

$$-19.54 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}_{\texttt{Intercept}} + 0.83 \begin{bmatrix} 32 \\ 33 \\ 32 \\ 33 \\ 26 \end{bmatrix}_{\texttt{age}} + 6.48 \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}_{\texttt{sexM}} = \begin{bmatrix} 13.65 \\ 8.00 \\ 13.65 \\ 14.48 \\ 8.64 \end{bmatrix}_{\texttt{fitted}}$$

Model Vectors and Redundancy

Vectors in the model matrix cannot be redundant. Vectors are redundant when one vector can be written as a *linear combination of the others*.

One common source of redundancy is when *all* indicator vectors for categorical variables are included in the model along with an intercept.

Consider the two sex indicator variables

$$1\begin{bmatrix}0\\1\\0\\0\\0\end{bmatrix}_{\texttt{sexF}} + 1\begin{bmatrix}1\\0\\1\\1\\1\end{bmatrix}_{\texttt{sexM}} = \begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}_{\texttt{Intercept}}$$

The intercept vector is just a linear combination of the two sex indicator vectors.

To fit a model that includes both indicators for sex, the intercept would need to be dropped from the model.

To drop the intercept, we include `-1` in the formula of the `lm()` function

```
> lm(wage ~ age + sex - 1, data = toy)

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
age    0.8346     0.4083   2.044    0.178
sexF -19.5402    13.6620  -1.430    0.289
sexM -13.0600    12.6054  -1.036    0.409
```

This results in the coefficients for the categorical predictors having different interpretations.

```
> model.matrix(lm.1)

  (Intercept) age sexM
1           1  32    1
2           1  33    0
3           1  32    1
4           1  33    1
5           1  26    1
```

The **model matrix** or **design matrix** can be obtained by inputting the name of a fitted model to the `model.matrix()` function

Geometry and Statistical Models

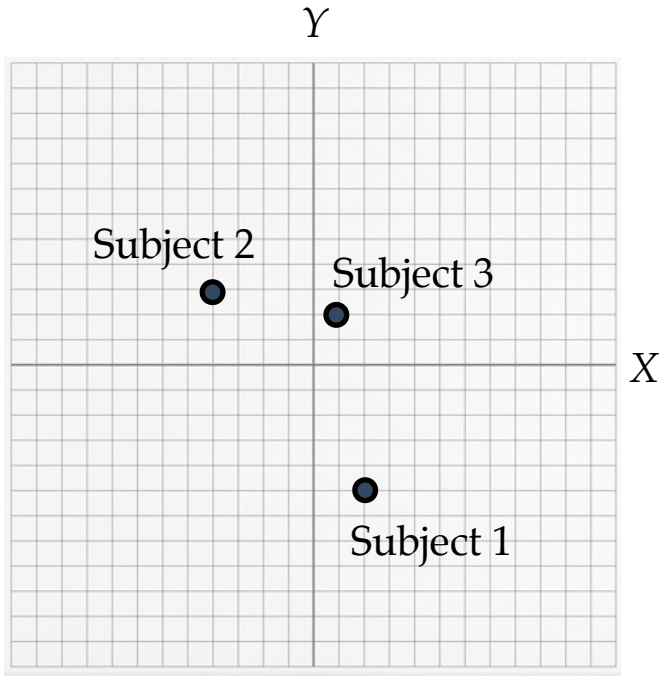
It is possible to compute all statistical calculations for regression models using only a ruler and protractor.

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

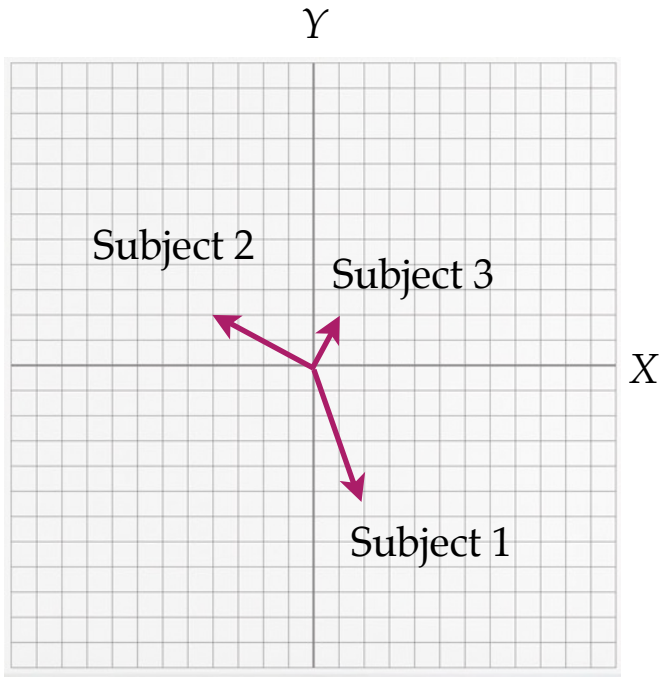
We typically plot multivariate data in variable space. In variable space, the **variables are represented as axes** and the **subjects are represented as points**, which are plotted based on their values for the variables.

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.

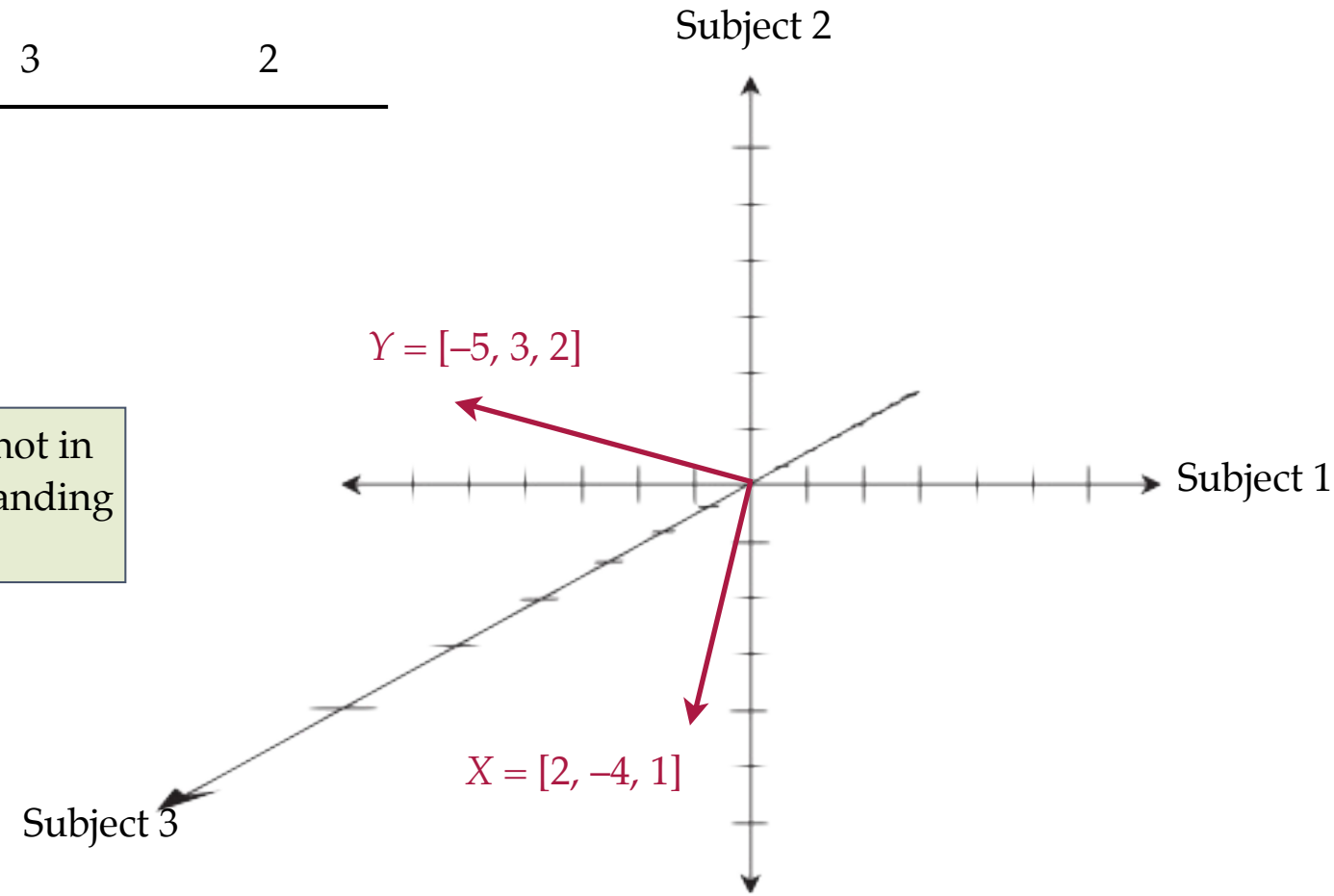
Subject	X	Y
Subject 1	2	–5
Subject 2	–4	3
Subject 3	1	2



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

Ithaca

Odysseus'
ship

Zephyros

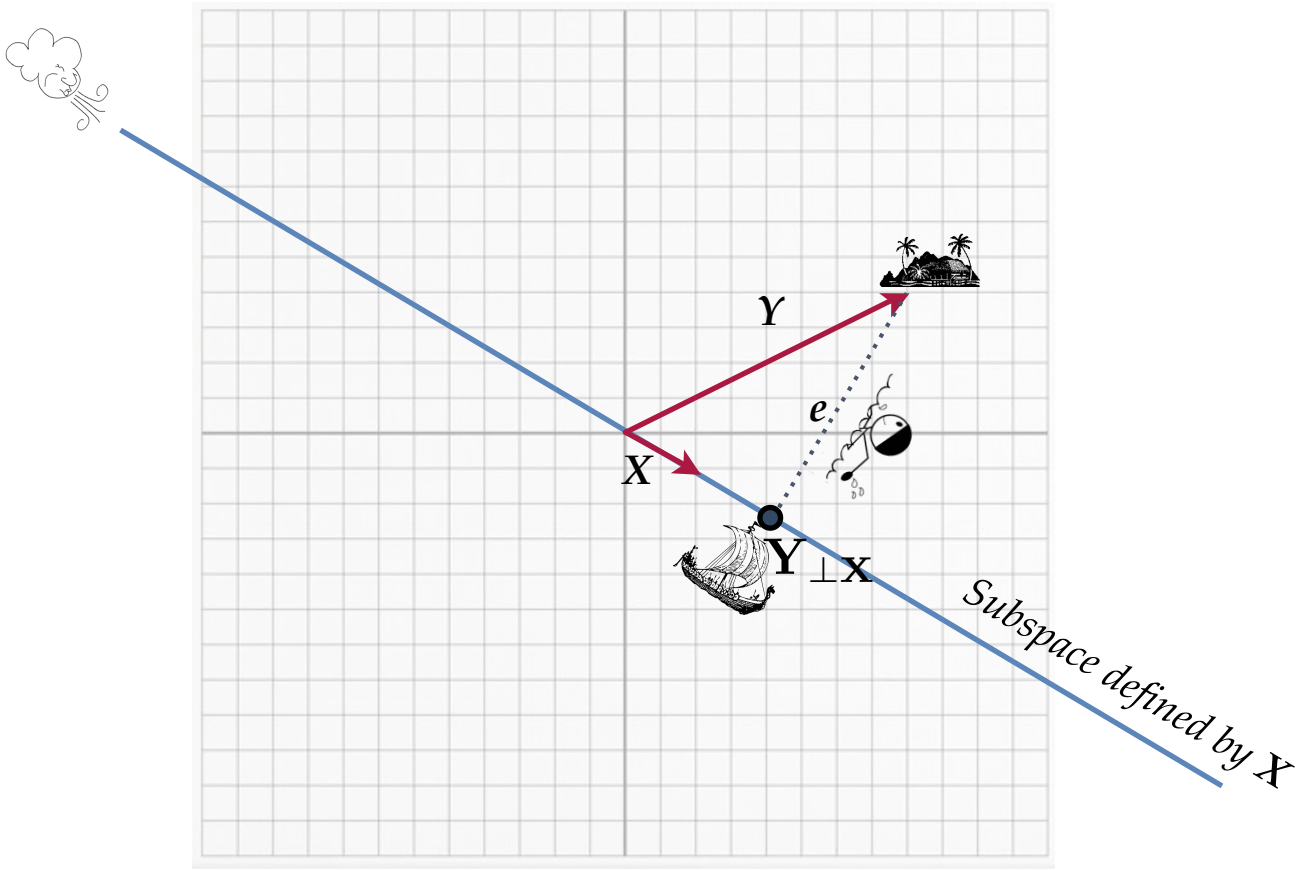
Aeolus
Island

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

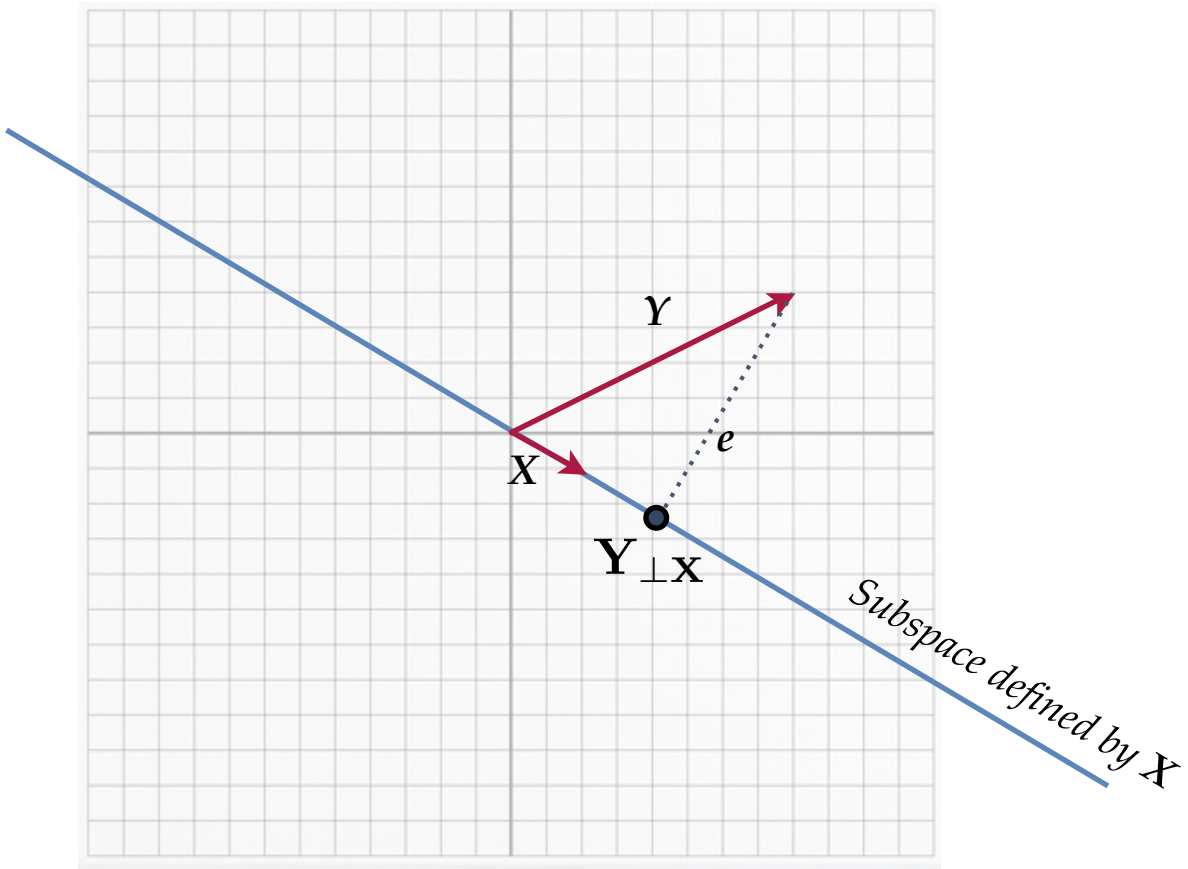
Odysseus wants to go home to Ithaca.

*Metaphor provided by Daniel Kaplan, Macalester

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.



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



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).

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

This is the vector of fitted values.

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

The residual is the vector between the point and the goal vector Y.

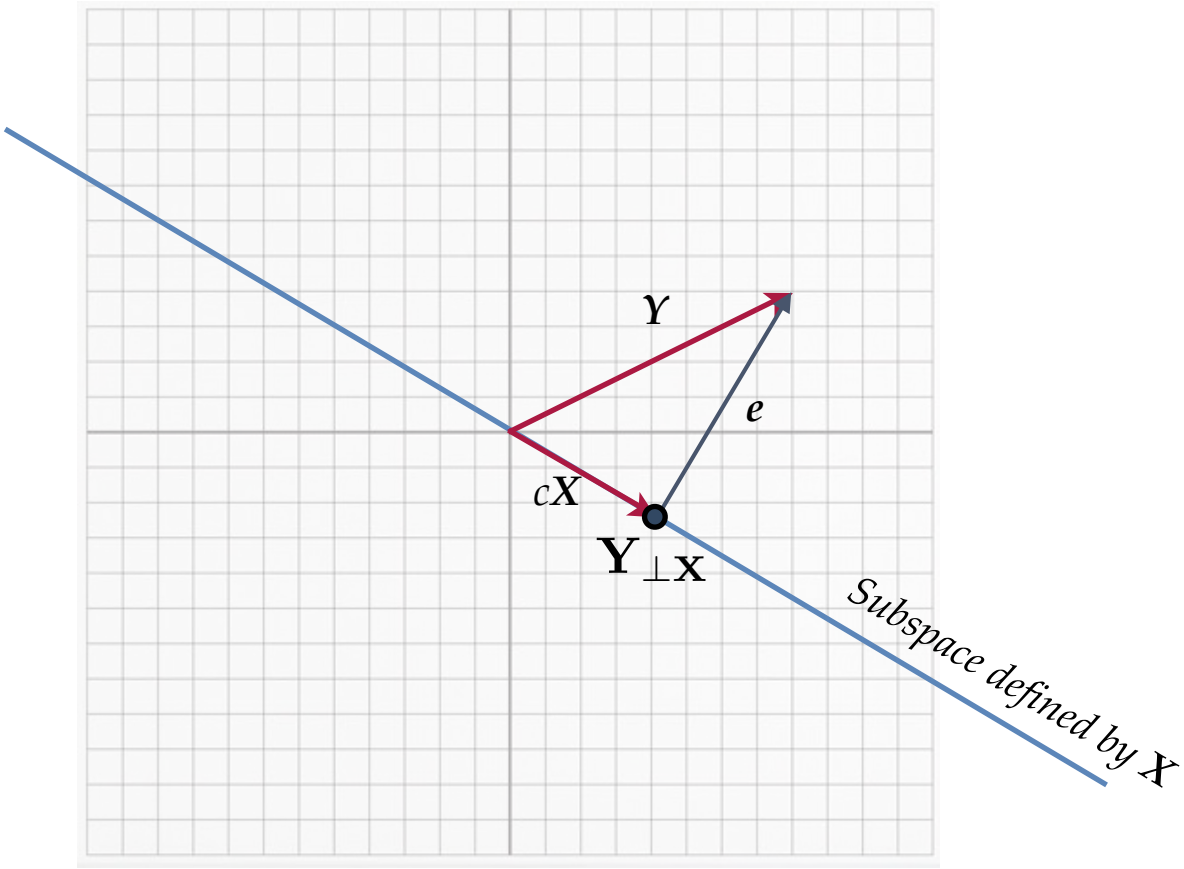
Since

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

Computing the residual vector is simply vector subtraction

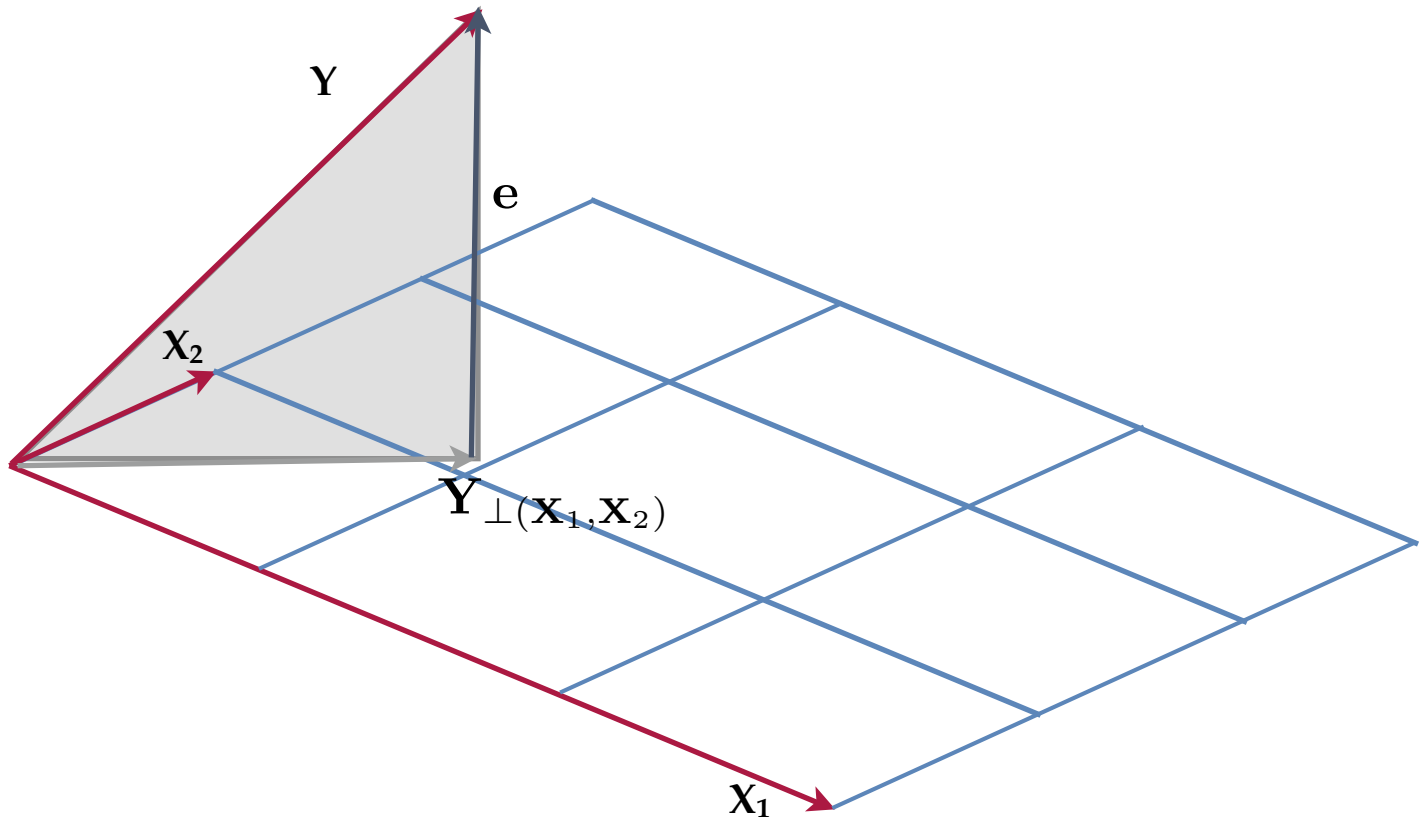
$$\mathbf{e} = \mathbf{Y} - c\mathbf{X}$$

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

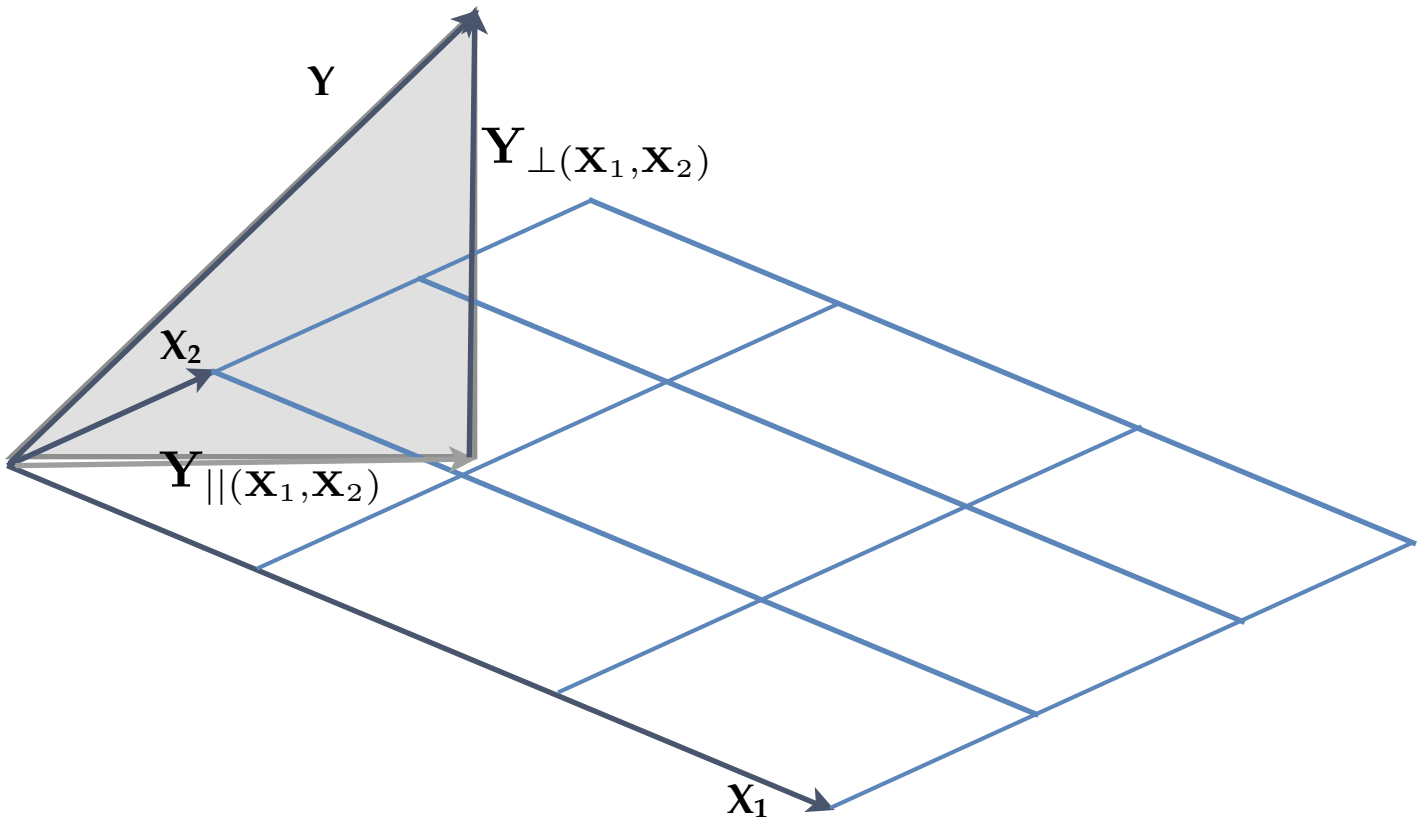


Multiple Predictors

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, $\bar{y} = 3,\ \hat{\sigma}_y = 2.83$

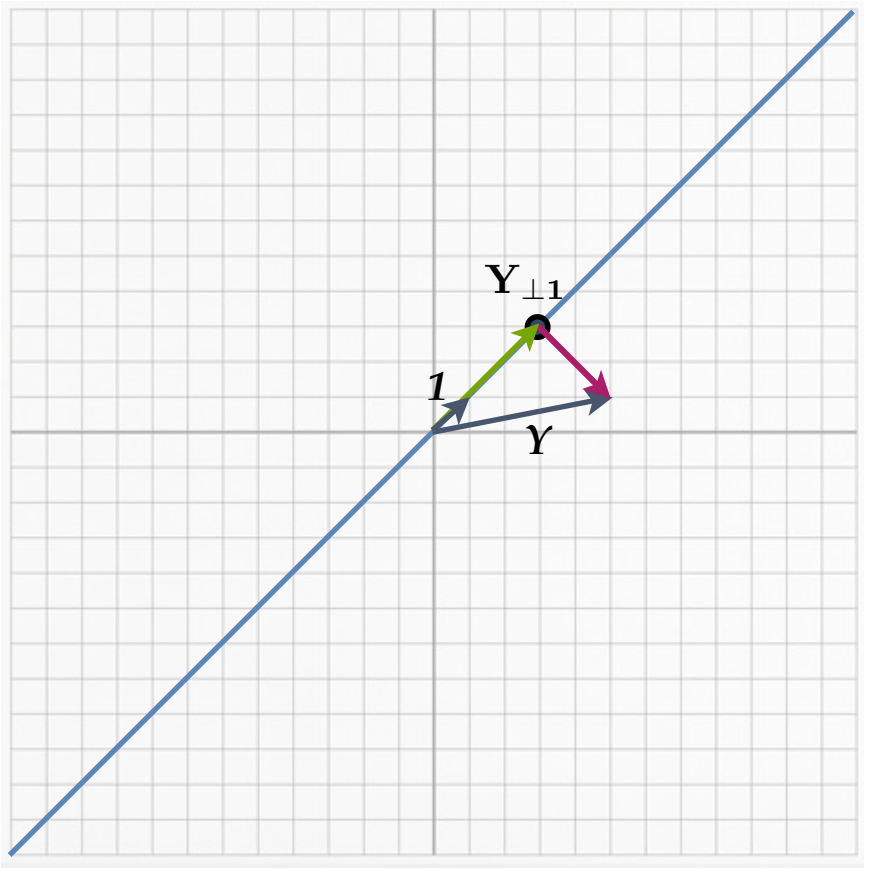
The intercept vector is a vector of ones.

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

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

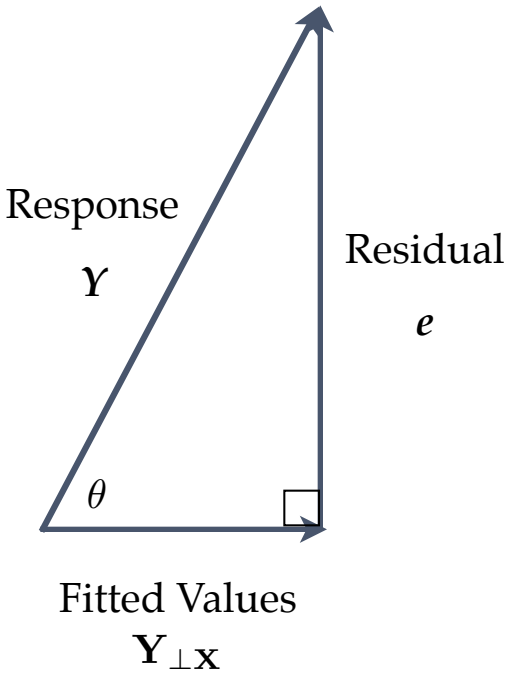
$$\mathbf{Y} - \mathbf{Y}_{\perp \mathbf{1}} = \begin{bmatrix} 5 \\ 1 \end{bmatrix} - \begin{bmatrix} 3 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ -2 \end{bmatrix}$$

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

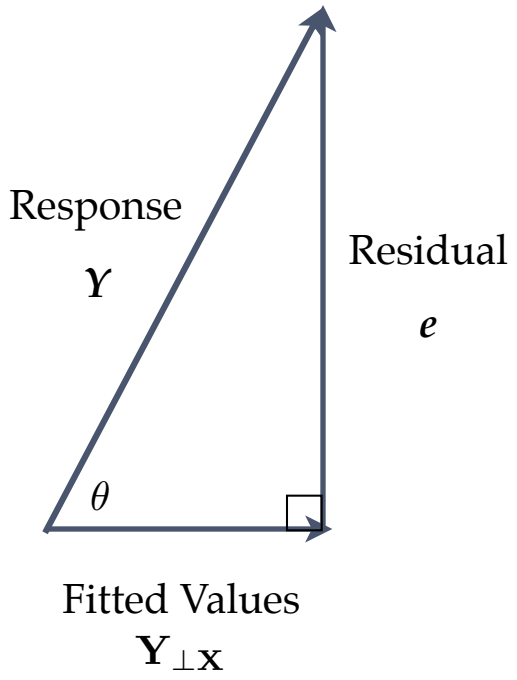
Since the model triangle is a right triangle, the Pythagorean Theorem relates the lengths of the three sides as

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

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

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

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

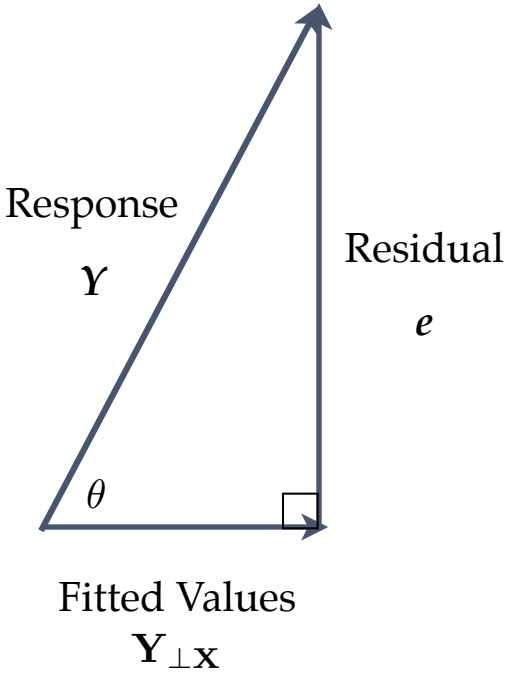


Correlation

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...

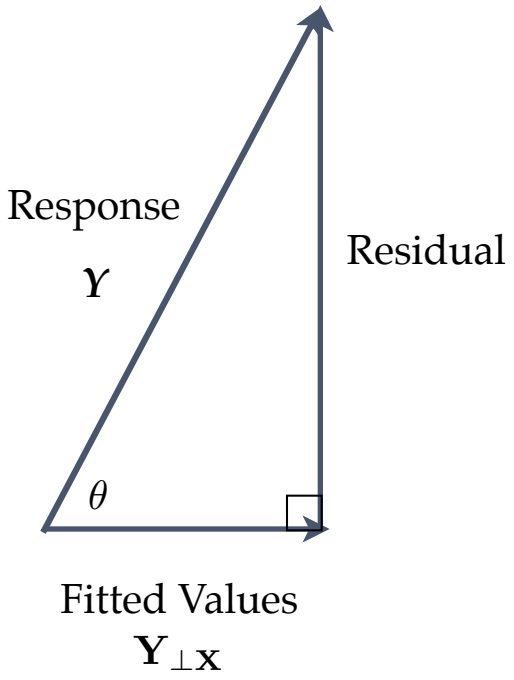
$$\left[\cos\theta\right]^2 = \frac{\mathbf{Y}_{\perp\mathbf{x}} \bullet \mathbf{Y}_{\perp\mathbf{x}}}{\mathbf{Y} \bullet \mathbf{Y}}$$

The numerator and denominator are both sum of squares!

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

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

$$\cos\theta = r$$



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.