A Primer on Vector Geometry

Measurement, Scaling, and Dimensional Analysis
2019 ICPSR Summer Program
Prof. Adam M. Enders

Housekeeping

- Blalock Lecture series: "The Promise of Nested Models for Critical Studies of Race and Racism"
  - Who: Abigail Sewell, Sociology, Emory University
  - When/Where: 7–9 PM, Angell Hall Auditorium D
  - July 9: "Creating Measures of Supraindividual Racism"
  - July 10: "Evaluating the Population Risks of Supraindividual Racism"
  - July 11: "Quantifying Intersectionality and Mixed Effects Models"
- Probably have lab covering PCA and factor analysis on Friday...

A Hypothetical Dataset

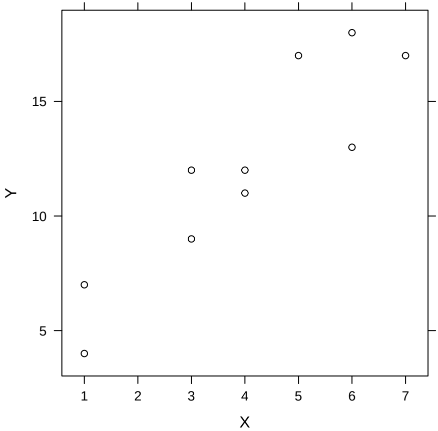
10 observations on two variables X and Y

i	X	Y
1	1	4
2	1	7
3	3	9
4	3	12
5	4	11
6	4	12
7	5	17
8	6	13
9	6	18
10	7	17

How might we represent this information visually?

A Scatterplot of Subjects

The obvious choice is a two-dimensional scatterplot where the variables are represented by the two axes and subjects are represent by points



Mean Centering

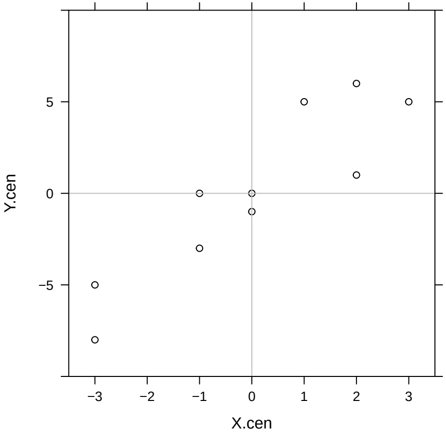
Though the location of the origin of the plot doesn't matter if we only care about the individual values of X and Y, our life will be easier in multivariate analyses if we shift the center of the plot to the origin (0, 0)

We can do this by subtracting the mean of each variable from every score: $x_i = X_i - \bar{X}$, $y_i = Y_i - \bar{Y}$

i	x	y
1	-3	-8
2	-3	-5
3	-1	-3
4	-1	0
5	0	-1
6	0	0
7	1	5
8	2	1
9	2	6
10	3	5

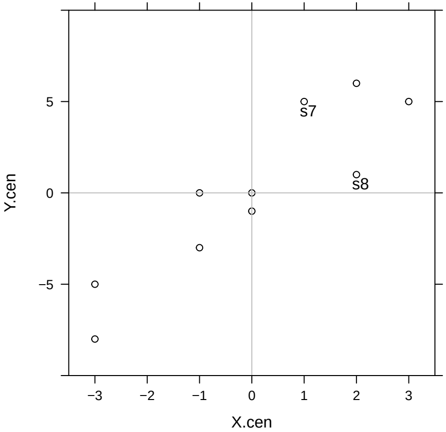
A (Mean-Centered) Scatterplot of Subjects

A two-dimensional depiction of the mean-centered data



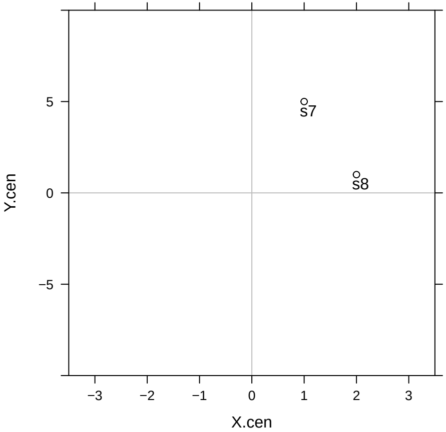
A Scatterplot of Subjects

Typical scatterplots place emphasis on the observations



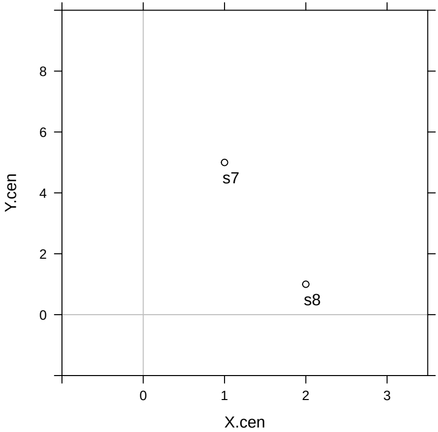
A Scatterplot of Subjects

Typical scatterplots place emphasis on the observations



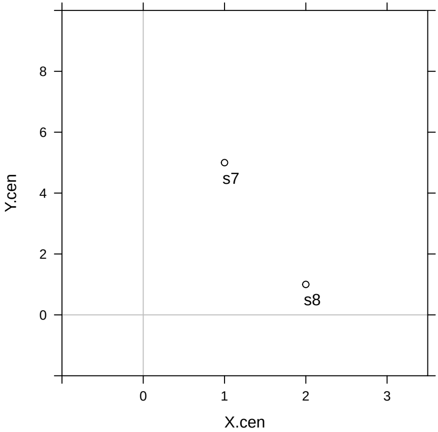
A Scatterplot of Subjects

Typical scatterplots place emphasis on the observations



A Scatterplot of Subjects

Typical scatterplots place emphasis on the observations

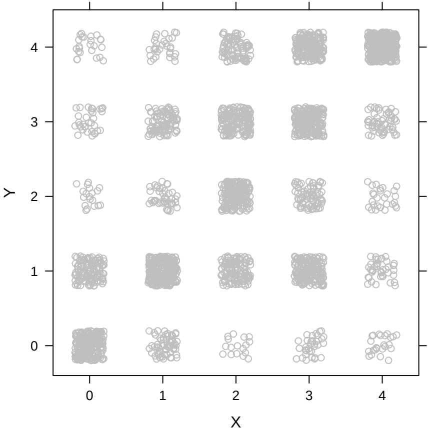


Do we really care about subject 7 and 8, though? Or, are we really interested in the relationships between the variables?

A Scatterplot of Subjects

It's hard to get much out of a scatterplot of "real" data

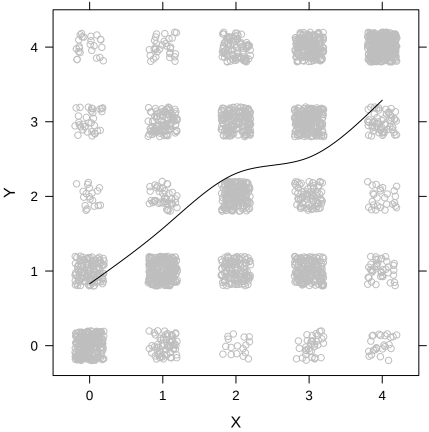
We're usually interested in relationship between variables – what's the correlation here?



A Scatterplot of Subjects

It's hard to get much out of a scatterplot of "real" data

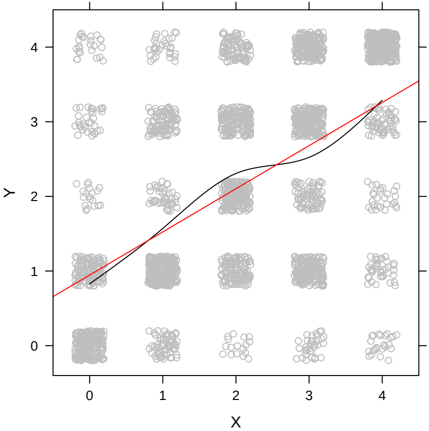
We're usually interested in relationship between variables – what's the correlation here?



A Scatterplot of Subjects

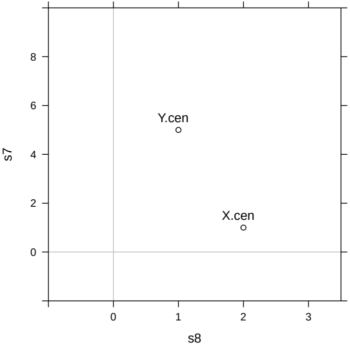
It's hard to get much out of a scatterplot of "real" data

We're usually interested in relationship between variables – what's the correlation here?



Abandoning Subjects for Variables

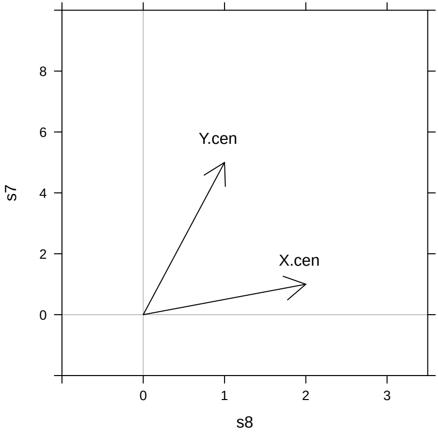
What if we made the subjects (row objects) the axes, so that variables are represented by points?



This is a scatterplot of subject 7 vs. subject 8, with the centered Y and X variables represented as points

Abandoning Subjects for Variables

Those points are actually terminal points of **vectors**



What's a Vector?

- Definition: a directed line segment emanating from the origin (0, 0) and terminating at the coordinates of the data point/observation
- Retains all of the information as the coordinates as long as it maintains its length and direction (angle)
- Usually helpful to work with vectors of unit length (1, that is)
  - We'll see how that's useful when we get to particular analyses, but don't need to worry too much about it for now

Variable Space and *n*-Space

- We are used to representing data in variable space; that is, we locate observations in a space where variables represent the coordinate axes
- Can also represent variables in subject space, called "*n*-space"
- We are going to work with vectors that have centered values:

$$\vec{x} = \left[x_1 - \bar{x}, x_2 - \bar{x}, \ldots x_n - \bar{x}\right]$$

- Which means that:

$$\begin{aligned}\vec{x} \cdot \vec{x} &= \left[(x_1 - \bar{x})(x_1 - \bar{x}) + (x_2 - \bar{x})(x_2 - \bar{x})\ldots + (x_n - \bar{x})(x_n - \bar{x})\right] \\ &= (x_1 - \bar{x})^2 + (x_2 - \bar{x})\ldots + (x_n - \bar{x})\end{aligned}$$

$$\vec{x}^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 \longleftarrow \text{ length of the variable vector}$$

- The length of the vector tells us the sum of squares, or, the variance of the vector (which represents a variable)

Vector Length

- Can also calculate length using the Pythagorean Theorem
  - Square of the hypotenuse (the side opposite the right angle) is equal to the sum of the squares of the other two sides
- Length can be expressed like the following if the mean has already been subtracted:

$$|\vec{x}| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$$

- Remember the two vectors from our example: $\vec{y} = [1, 5]'$, and $\vec{x} = [2, 1]'$
- Their lengths are as follows:

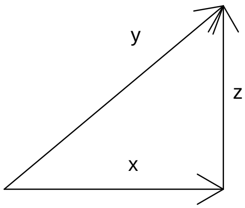
$$|\vec{x}| = \sqrt{2^2 + 1^2} = \sqrt{5} = 2.236$$
$$|\vec{y}| = \sqrt{1^2 + 5^2} = \sqrt{26} = 5.099$$

Relationships in a Right Triangle

- The Pythagorean Theorem holds the following:

$$|\vec{x}|^2 + |\vec{z}|^2 = |\vec{y}|^2$$

- Thus, when the lengths of any two vectors are known, we can easily solve for the length of the third vector



Arithmetical Operations: Scalar Multiplication

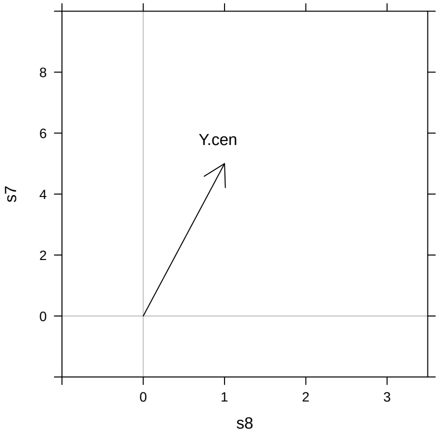
- Scalar multiplication: changes the length of the vector, but doesn't change the orientation
- If k is greater than 1 the length increases, if it is less than 1 it shortens (if it is 1, nothing changes)
- Multiplication by a negative constant k will change the direction (its now "opposite"), but not the orientation

$$\vec{x} = \left[x_1, x_2\right]$$
$$k\vec{x} = \left[x_1 k, x_2 k\right]$$

- Scalar multiplication of a vector *generates a subspace* of x because any constant will leave the vector on the original collinear line
- Subspace of $\vec{x}, \vec{y}$ is plane, p

Arithmetical Operations: Scalar Multiplication

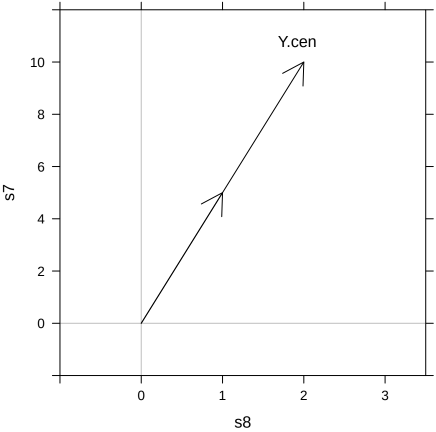
Focus on $\vec{y} = [1, 5]'$



Arithmetical Operations: Scalar Multiplication

$2\vec{y} = [2 \times 1, 2 \times 5]' = [2, 10]'$

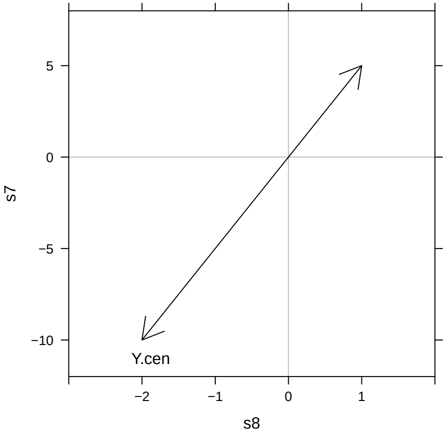
Multiplying by 2 doubles the length of the vector in the current direction



Arithmetical Operations: Scalar Multiplication

$$-2\vec{y} = [-2 \times 1, -2 \times 5]' = [-2, -10]'$$

Same increase in length, just reversed directions



Arithmetical Operations: Addition

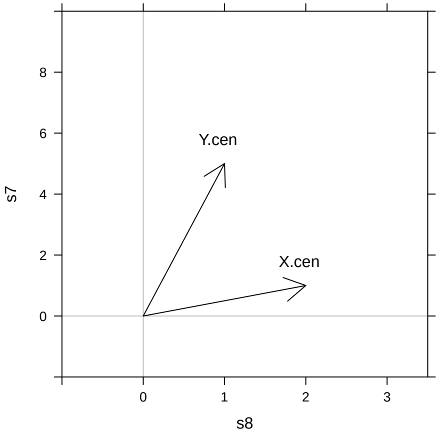
- Addition:

$$\vec{x} = \left[x_1, x_2\right]$$
$$\vec{y} = \left[y_1, y_2\right]$$
$$\vec{x} + \vec{y} = \left[x_1 + y_1, x_2 + y_2\right]$$

- Geometrically, the sum of two vectors is produced by moving one of the vectors to the end of the other and drawing the sum as a vector from the start of the first vector to the end of the second

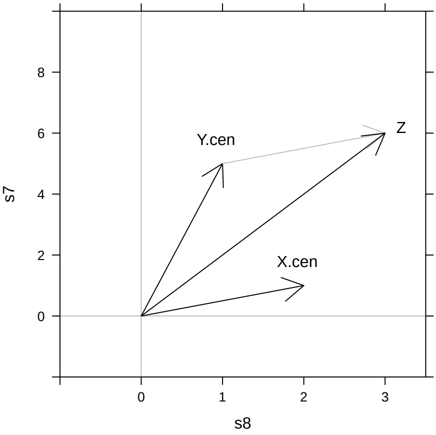
Arithmetical Operations: Addition

$$\vec{z} = \vec{x} + \vec{y} = [2+1, 1+5]' = [6,3]'$$



Arithmetical Operations: Addition

$$\vec{z} = \vec{x} + \vec{y} = [2 + 1, 1 + 5]' = [6, 3]'$$



Arithmetical Operations: Subtraction

- Subtraction works just like addition:

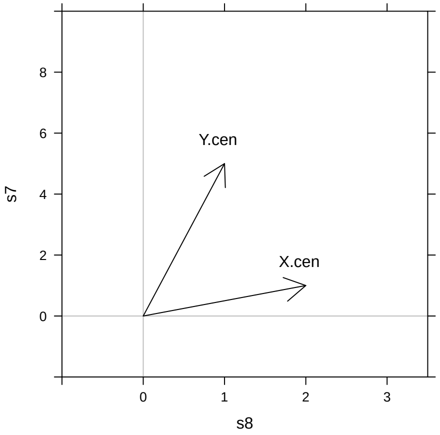
$$\vec{x} = \left[x_1, x_2\right]$$
$$\vec{y} = \left[y_1, y_2\right]$$
$$\vec{x} - \vec{y} = \left[x_1 - y_1, x_2 - y_2\right]$$

$\vec{x} - \vec{y}$ is also equal to $\vec{x} + (-1 \times \vec{y})$

- Geometrically, subtraction corresponds to going to the end of the first vector and moving in a direction opposite to that of the second vector

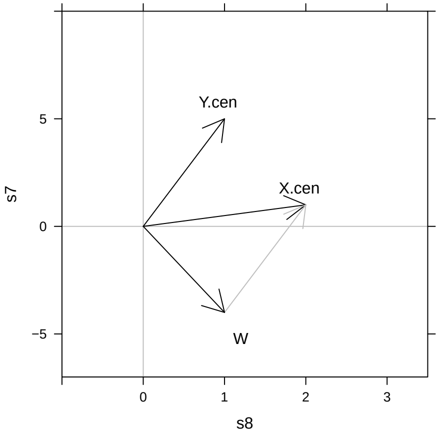
Arithmetical Operations: Subtraction

$$\vec{w} = \vec{x} - \vec{y} = [2 - 1, 1 - 5]' = [1, -4]'$$



Arithmetical Operations: Subtraction

$$\vec{w} = \vec{x} - \vec{y} = [2 - 1, 1 - 5]' = [1, -4]'$$



Arithmetical Operations: Linear Combinations

- A linear combination involves both addition and scalar multiplication
- Regression equations, and as we will see shortly, principal components analysis, are linear combinations

$$\vec{z} = b_x\vec{x} + b_y\vec{y}$$

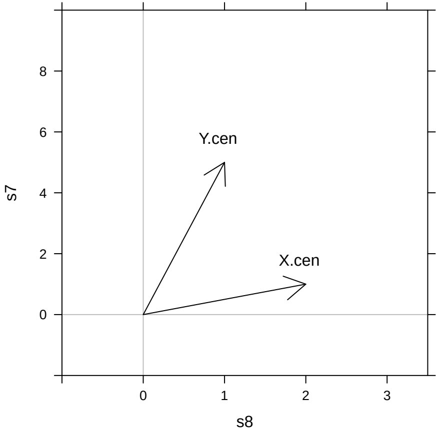
- Geometrically, the above equation entails moving along $\vec{x}$ for a distance of b_x times its length, then turning in the direction of $\vec{y}$ for for b_y times its length

Arithmetical Operations: Linear Combinations

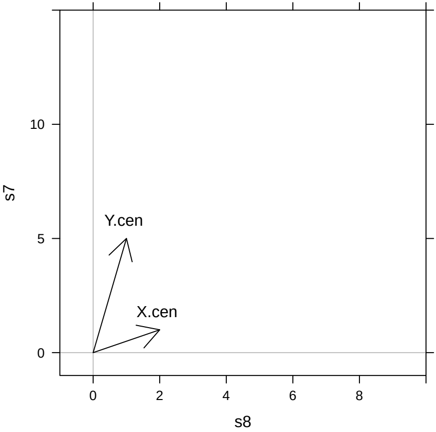
Continuing with our hypothetical vectors $\vec{x}$ and $\vec{y}$ from before...

$$\begin{aligned}
\vec{z} &= b_x\vec{x} + b_y\vec{y} \\
&= 3\vec{x} + 2\vec{y} \\
&= 3\begin{bmatrix} 2 \\ 1 \end{bmatrix} + 2\begin{bmatrix} 1 \\ 5 \end{bmatrix} \\
&= \begin{bmatrix} 6 \\ 3 \end{bmatrix} + \begin{bmatrix} 2 \\ 10 \end{bmatrix} \\
&= \begin{bmatrix} 8 \\ 13 \end{bmatrix}
\end{aligned}$$

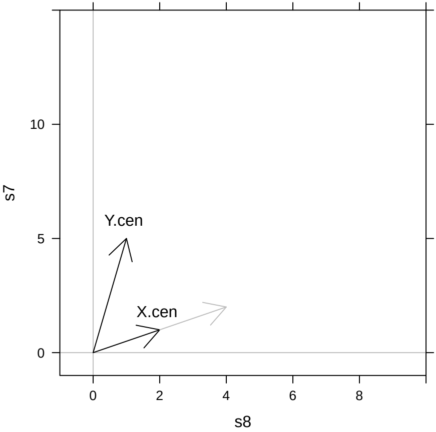
Arithmetical Operations: Linear Combinations



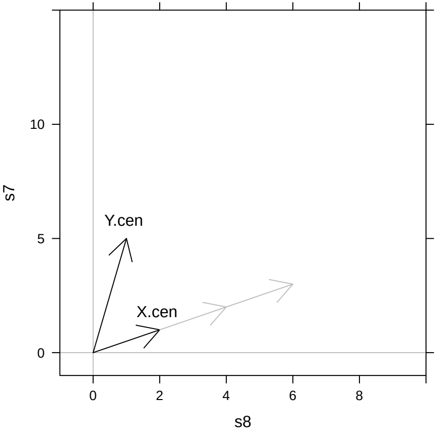
Arithmetical Operations: Linear Combinations



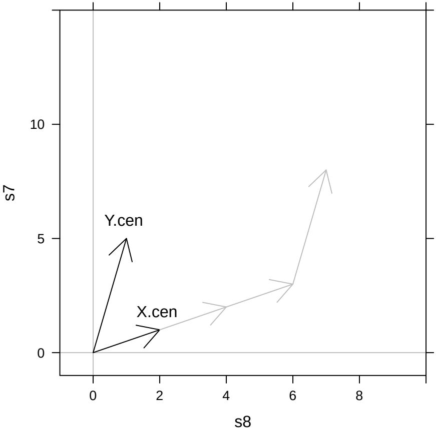
Arithmetical Operations: Linear Combinations



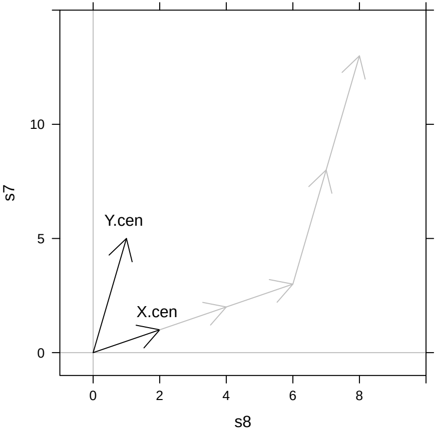
Arithmetical Operations: Linear Combinations



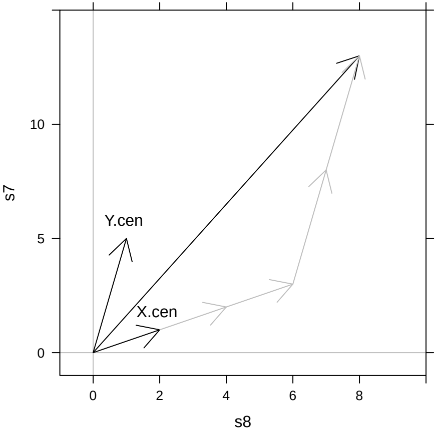
Arithmetical Operations: Linear Combinations



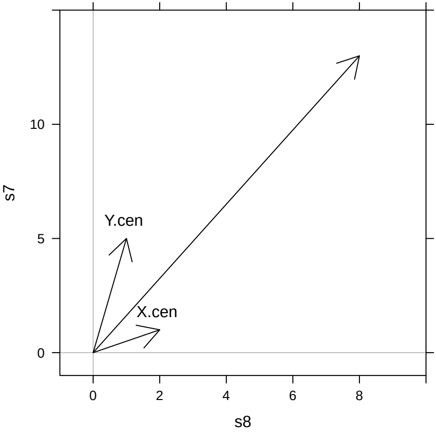
Arithmetical Operations: Linear Combinations



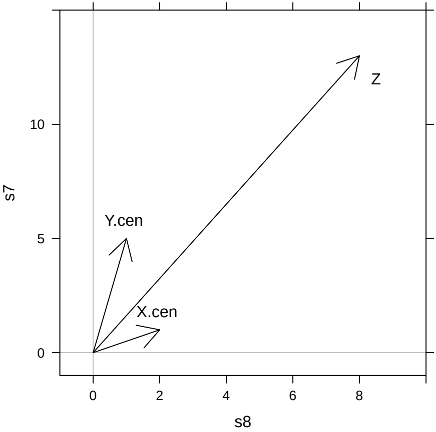
Arithmetical Operations: Linear Combinations



Arithmetical Operations: Linear Combinations



Arithmetical Operations: Linear Combinations



Arithmetical Operations: Scalar Product

- Scalar product: scalar number formed by sum of products of corresponding elements of $\vec{x}$ and $\vec{y}$
  - Also called the "dot product" because the operation is often represented $\vec{x} \cdot \vec{y}$

$$\vec{x}\vec{y} = x_1 y_1 + x_2 y_2 ... + x_m y_m$$

or

$$\vec{x}\vec{y} = \sum^{m} x_m y_m$$

  - Scalar product of the vector with itself is equal to the length of that vector
  - Length of $\vec{x}$ (which is $\vec{x} \cdot \vec{x}$): $|\vec{x}|$ or $||\vec{x}||$
  - Scalar product of two vectors is also equal to the cosine of the angle formed between the vectors: $\vec{x}\vec{y} = |\vec{x}|\vec{y}|cos\theta_{\vec{x}\vec{y}}$

Arithmetical Operations: Correlation

- Correlation: corresponds to the cosine of the angle between two vectors (ie, $|\vec{x}||\vec{y}|cos\theta_{\vec{x}\vec{y}}$)

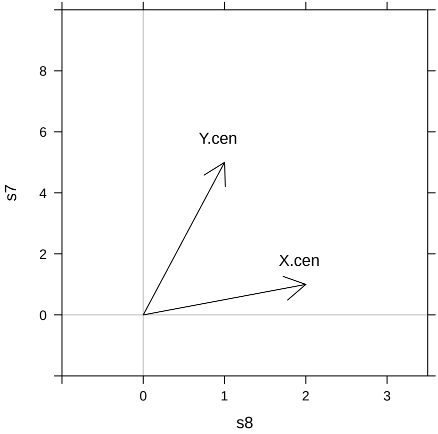
$$\vec{x}\cdot\vec{y} = |\vec{x}||\vec{y}|cos\theta_{\vec{x}\vec{y}}$$
$$\frac{\vec{x}\cdot\vec{y}}{|\vec{x}||\vec{y}|} = cos\theta_{\vec{x}\vec{y}}$$
$$cos\theta_{\vec{x}\vec{y}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

  - Orthogonal vectors (90 degree/right angle) are uncorrelated ($cos90 = 0$)
    - Furthermore, two orthogonal vectors will have a dot product of 0 ($x \perp y = x \cdot y = 0$)
  - Two vectors forming an acute angle (between 90 and 0 degrees) are positively correlated
  - Two vectors forming an obtuse angle (between 90 and 180 degrees) are negatively correlated

Arithmetical Operations: Correlation

$$cos\theta_{\vec{x}\vec{y}} = \frac{\vec{x}\cdot\vec{y}}{|\vec{x}||\vec{y}|} = \frac{7}{5.099 \times 2.236} = 0.614, \text{ about } 52^{\circ}$$

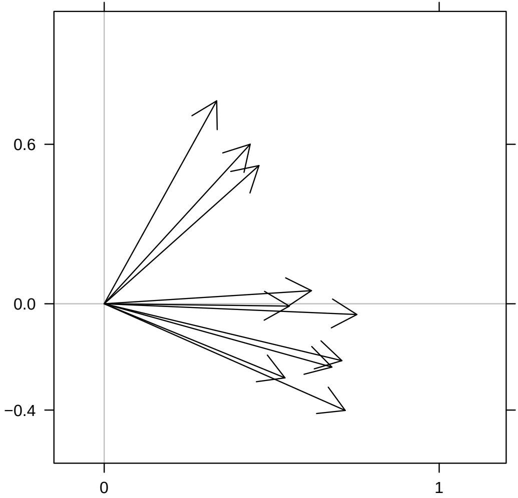
Provides a much "cleaner" visualization of the relationship between variables



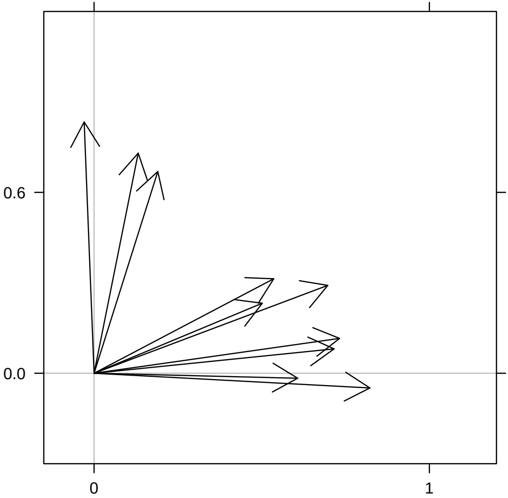
Vector Rotation

- A pair of vectors can be rotated by some angle θ by rigidly turning them through this angle without altering the angle between them
- Rotating a set of vectors changes their orientation to the coordinate system you're using, but not the angular relationships between vectors or lengths of the vectors
- Essentially, the process is akin to looking at the vectors from a new viewpoint
- Can think of rotation in terms of rotating the coordinate axes, rather than the vectors
- Either way, the original vectors are multiplied by a rotation vector comprised of direction cosines
  - The resultant vector contains the coordinates in the rotated space

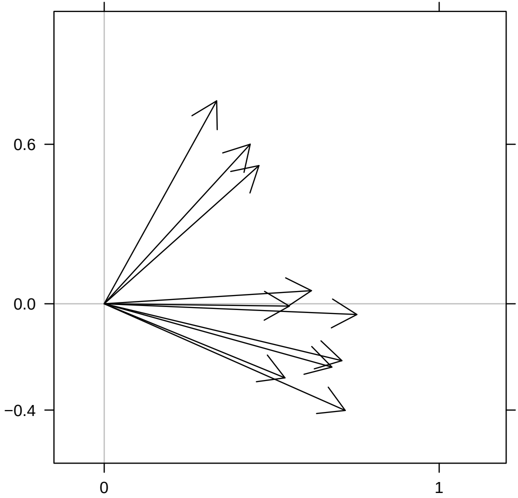
Vector Rotation



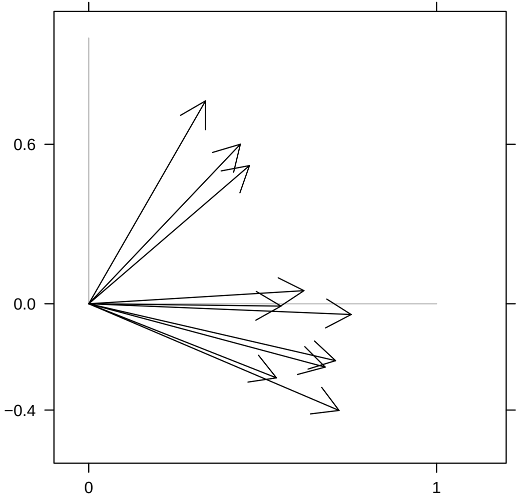
Vector Rotation



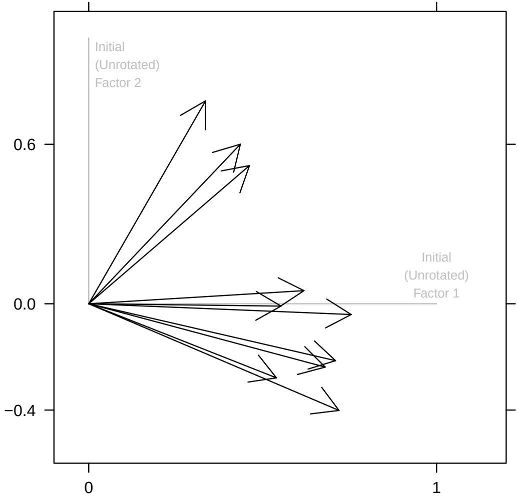
Vector Rotation



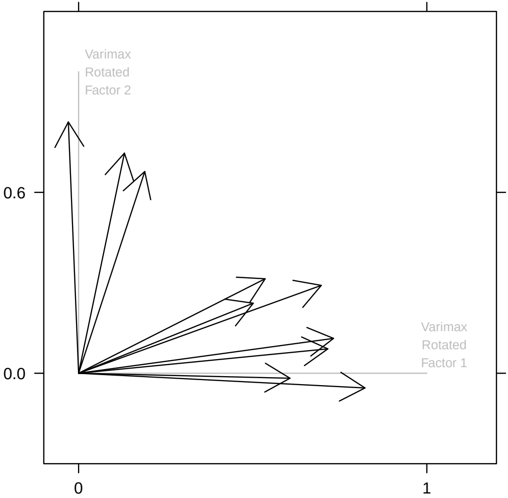
Vector Rotation



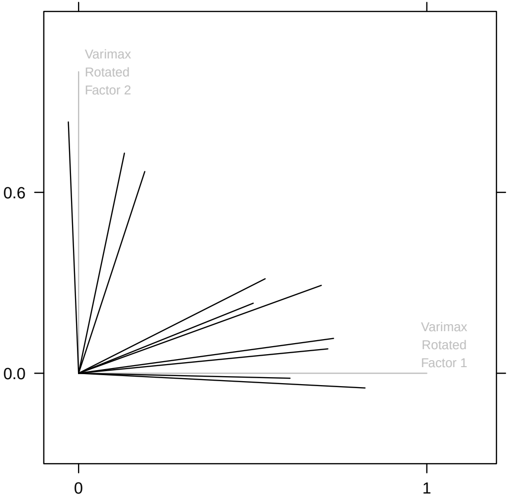
Vector Rotation



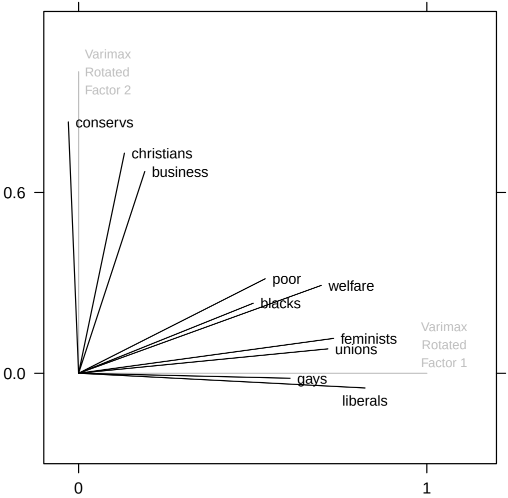
Vector Rotation



Vector Rotation



Vector Rotation



Regression with Vector Geometry

- The standard prediction equation for OLS regression:

$$\hat{Y} = a + bX$$

- Let's think of $\hat{Y}$ and X in terms of vectors, $\vec{\hat{y}}$ and $\vec{x}$
- Furthermore, let's center the data (per usual in multivariate stats) so we can drop the intercept and simplify:

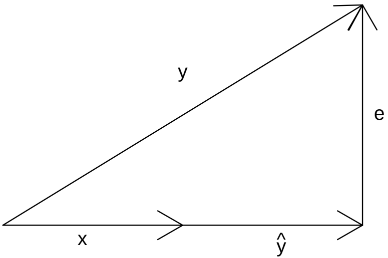
$$\vec{\hat{y}} = b\vec{x}$$

- Since $\vec{\hat{y}}$ is a scalar multiple of $\vec{x}$ (via b), we know that $\vec{\hat{y}}$ is collinear to $\vec{x}$
- Furthermore, unless $\vec{x}$ and $\vec{y}$ are perfectly collinear (correlated), which would make the problem trivial, $\vec{y}$ points in a different direction than $\vec{x}$
- Objective: find the value b that would make $\vec{y}$ and $\vec{\hat{y}}$ as similar as possible, or:

$$\vec{e} = \vec{y} - \vec{\hat{y}}$$

Regression with Vector Geometry

$\vec{e}$ must be orthogonal – uncorrelated – with $\vec{\hat{y}}$, which is collinear with $\vec{x}$



Regression with Vector Geometry

- The orthogonality of $\vec{e}$ and $\vec{x}/\vec{\hat{y}}$ let's us calculate b
- When b is chosen correctly, the dot product $\vec{x} \cdot \vec{e} = 0$
- We can use the previous equations substituting $\vec{e}$ with $\vec{y} - b\vec{x}$:

$$
\begin{aligned}
\vec{x} \cdot (\vec{y} - b\vec{x}) &= 0 \\
\vec{x} \cdot \vec{y} - b(\vec{x} \cdot \vec{x}) &= 0 \\
\vec{x} \cdot \vec{y} - b(|\vec{x}|^2) &= 0 \\
b &= \frac{\vec{x} \cdot \vec{y}}{|\vec{x}|^2}
\end{aligned}
$$

- This is the vector geometric expression of the algebraic bivariate regression coefficient:

$$
b = \frac{\sum x_i y_i}{\sum x_i^2}
$$

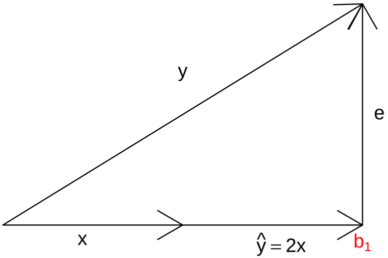
Bivariate Regression: An Example

- Take the following vectors, for example:
  - $\vec{x} = [-3, -3, -1, -1, 0, 0, 1, 2, 2, 3]'$
  - $\vec{y} = [-8, -5, -3, 0, -1, 0, 5, 1, 6, 5]'$
- $|\vec{x}|^2 = 38$, $\vec{x} \cdot \vec{y} = 76$

$$b = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}|^2} = \frac{76}{38} = 2.00$$

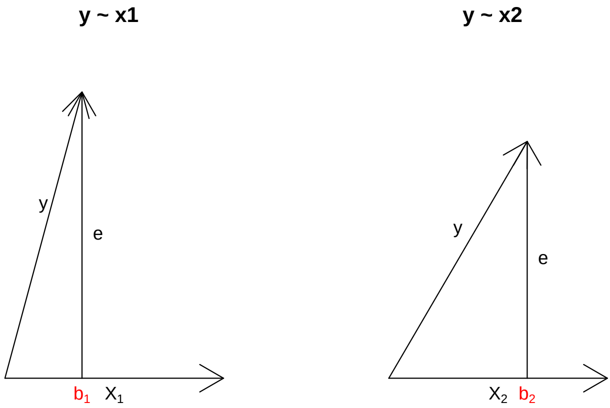
Bivariate Regression: An Example

$$b = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}|^2} = \frac{76}{38} = 2.00$$



Goodness of Fit

The more highly correlated $\vec{x}$ is with $\vec{y}$, the smaller the angle between $\vec{x}$ and $\vec{y}$, the greater the fit (in terms of b and R^2):



Goodness of Fit

- We know that the cosine of the angle between two vectors corresponds to the correlation
- We know that the squared correlation between x and y in a bivariate regression is the R^2
  - Relatedly, we know that the correlation between $\hat{y}$ and y, also gives the R^2
  - And, $\vec{\hat{y}}$ is collinear with $\vec{x}$
- Thus, the R^2 is the square of $cos\theta_{\vec{x}\vec{y}}$
- In this example:

$$cos\theta_{xy} = \frac{x \cdot y}{|x||y|} = 0.903 \longrightarrow R^2 = 0.817$$

More Than One Independent Variable

- The bivariate case generalizes to as many dimensions as there are variables in the model (assuming none are perfectly collinear)
- Can't really picture a case with more than three variables (including the DV), but the math works
- In three dimensions (2 IVs and a DV):
  - Think of $\vec{x}_1$ and $\vec{x}_2$ as occupying same subspace, a plane called the "regression space"
  - Since $\vec{\hat{y}}$ is a linear combination of $b_1\vec{x}_1 + b_2\vec{x}_2$, $\vec{\hat{y}}$ also resides in the same "regression" subspace as $\vec{x}_1$ and $\vec{x}_2$
  - $\vec{e}$ remains orthogonal to $\vec{x}_1$, $\vec{x}_2$, and $\vec{\hat{y}}$, but occupies another subspace called the "error space"
  - Still want to pick b_1 and b_2 so that the tip of $\vec{\hat{y}}$ is directly under $\vec{y}$

More Than One Independent Variable

- Can no longer calculate R^2 as the square of $cos\theta_{\vec{x}\vec{y}}$
- Rather, we calculate goodness of fit as the length of $\vec{\hat{y}}$, $|\vec{\hat{y}}|$

$$R = cos\theta_{\vec{\hat{y}}\vec{y}} = \frac{|\vec{\hat{y}}|}{|\vec{y}|}$$

- The longer $|\vec{\hat{y}}|$, the better the fit
- Standardized variables:
  - Can standardize variables by subtracting their mean and dividing by their standard deviation
  - In a geometric sense, this has the effect of standardizing the length of all variables to 1, or "unit length"
  - Then, we can interpret regression coefficients as "a one standard deviation increase in x leads to a ____ standard deviation decrease in y, on average"