Linear algebra for statistics

- Linear algebra is the math of vectors and matrices.
- In statistics, the main purpose of linear algebra is to organize data and write down the manipulations we want to do to them.
- A **vector** of length n is also called an n-**tuple**, or an **ordered sequence** of length n.
- We can suppose that each data point is a **real number**. We write $\mathcal{R}$ for the set of real numbers, and $\mathcal{R}^n$ for the set of vectors of n real numbers.
- Write the US life expectancy at birth for 2011 to 2015 as $y = (y_1, y_2, y_3, y_4, y_5) = (79.0, 79.1, 79.0, 79.0, 78.9)$.
- We see $y \in \mathcal{R}^5$. Numerical data can always be written as a vector in $\mathcal{R}^n$ where n is the number of datapoints. Categorical data can also be written as a vector in $\mathcal{R}^n$ by assigning a number for each category.

Question: You may or may not have seen vectors in other contexts. In physics, a vector is a quantity with magnitude and direction. How does that fit in with our definition?

Adding vectors and multiplying by a scalar: Example

- For a dataset, the **index** i of the **component** y_i of the vector y might correspond to a measurement on the ith member of a population, the outcome of the ith group in an experiment, or the ith observation out of a sequence of observations on a system. Generically, we will call i an **observational unit**, or just **unit**.
- We might want to add two quantities u_i and v_i for unit i.
- Using vector notation, if $u = (u_1, u_2, \ldots, u_n)$, $v = (v_1, v_2, \ldots, v_n)$ and $y = (y_1, y_2, \ldots, y_n)$ we define the **vector sum** $y = u + v$ to be the **componentwise sum** $y_i = u_i + v_i$, adding up the corresponding components for each unit.
- We might also want to rescale each component by the same factor. To change a measurement y_i in inches to a new measurement z_i in mm, we rescale with the **scalar** $\alpha = 25.4$. We want $z_i = \alpha y_i$ for each i. This is written in vector notation as **multplication of a vector by a scalar**, $z = \alpha y$.
- Keep track of whether each object is a scalar, a vector (what is its length?) or a matrix (what are its dimensions?).

Adding vectors and multiplying by a scalar: Example

An ecologist measures the pH of ten Michigan lakes at two points in the summer. Set up vector notation to describe her data. Write a vector calculation to find the average pH in each lake.

Adding vectors and multiplying by a scalar: Example

An ecologist measures the pH of ten Michigan lakes at two points in the summer. Set up vector notation to describe her data. Write a vector calculation to find the average pH in each lake.

First, set up notation.

Let x_i be the first pH measurement in lake i, for $i \in \{1, 2, \ldots, 10\}$.

Then, $x = (x_1, \ldots, x_{10})$ is the vector of the first pH measurement in each of the 10 lakes.

Let $y = (y_1, \ldots, y_{10})$ be the vector of second measurements.

Let $\mu = (\mu_1, \ldots, \mu_{10})$ be the average pH for each of the 10 lakes.

Adding vectors and multiplying by a scalar: Example

An ecologist measures the pH of ten Michigan lakes at two points in the summer. Set up vector notation to describe her data. Write a vector calculation to find the average pH in each lake.

First, set up notation.

Let x_i be the first pH measurement in lake i, for $i \in \{1, 2, \ldots, 10\}$.

Then, $x = (x_1, \ldots, x_{10})$ is the vector of the first pH measurement in each of the 10 lakes.

Let $y = (y_1, \ldots, y_{10})$ be the vector of second measurements.

Let $\mu = (\mu_1, \ldots, \mu_{10})$ be the average pH for each of the 10 lakes.

For each lake i, the mean is $\mu_i = \frac{1}{2}(x_i + y_i)$. In vector notation, this is $\mu = \frac{1}{2}(x + y)$.

Vectors and scalars in R

- We have seen in Chapter 1 that R has vectors. An R vector of length 1 could be called a scalar.
- You can check that R follows the usual mathematical rules of vector addition and multiplication by a a scalar.