



Some basic notation and background

Regression

Brian Caffo, PhD
Johns Hopkins Bloomberg School of Public Health

Some basic definitions

- In this module, we'll cover some basic definitions and notation used throughout the class.
- We will try to minimize the amount of mathematics required for this class.
- No calculus is required.

Notation for data

- We write X_1, X_2, \dots, X_n to describe n data points.
- As an example, consider the data set $\{1, 2, 5\}$ then
 - $X_1 = 1, X_2 = 2, X_3 = 5$ and $n = 3$.
- We often use a different letter than X , such as Y_1, \dots, Y_n .
- We will typically use Greek letters for things we don't know. Such as, μ is a mean that we'd like to estimate.
- We will use capital letters for conceptual values of the variables and lowercase letters for realized values.
 - So this way we can write $P(X_i > x)$.
 - X_i is a conceptual random variable.
 - x is a number that we plug into.

The empirical mean

- Define the empirical mean as

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

- Notice if we subtract the mean from data points, we get data that has mean 0. That is, if we define

$$\tilde{X}_i = X_i - \bar{X}.$$

$$\begin{aligned} \text{sum}(\tilde{X}_i) &= \text{sum}(X_i - \bar{X}) = \text{sum}(X_i) - \text{sum}(\bar{X}) = n\bar{X} - n * \bar{X} = 0 \\ &= n\bar{X} \text{ da } \bar{X} = \text{sum}(X_i) / n \end{aligned}$$

The the mean of the \tilde{X}_i is 0.

- This process is called "centering" the random variables.
- The mean is a measure of central tendency of the data.
- Recall from the previous lecture that the mean is the least squares solution for minimizing

$$\sum_{i=1}^n (X_i - \mu)^2$$

The empirical standard deviation and variance

- Define the empirical variance as

einfacher zum Rechnen:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right)$$

- The empirical standard deviation is defined as $S = \sqrt{S^2}$. Notice that the standard deviation has the same units as the data.
- The data defined by X_i/s have empirical standard deviation 1. This is called "scaling" the data.
- The empirical standard deviation is a measure of spread.
- Sometimes people divide by n rather than $n-1$ (the latter produces an unbiased estimate.)

nicht so wichtig, wenn n gross

Normalization

- The the data defined by

$$Z_i = \frac{X_i - \bar{X}}{s}$$

have empirical mean zero and empirical standard deviation 1.

- The process of centering then scaling the data is called "normalizing" the data.
- Normalized data are centered at 0 and have units equal to standard deviations of the original data.
- Example, a value of 2 from normalized data means that data point was two standard deviations larger than the mean.

The empirical covariance

- Consider now when we have pairs of data, (X_i, Y_i) .
- Their empirical covariance is

einfachere Formel fuer Rechnen:

$$\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n-1} \left(\sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y} \right)$$

X, Y : Vektoren
 Cov : eine Funktion

- Some people prefer to divide by n rather than $n - 1$ (the latter produces an unbiased estimate.)
- The correlation is defined is

$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{S_x S_y}$$

where S_x and S_y are the estimates of standard deviations for the X observations and Y observations, respectively.

Some facts about correlation

- $\text{Cor}(X, Y) = \text{Cor}(Y, X)$ symmetrisch
- $-1 \leq \text{Cor}(X, Y) \leq 1$
- $\text{Cor}(X, Y) = 1$ and $\text{Cor}(X, Y) = -1$ only when the X or Y observations fall perfectly on a positive or negative sloped line, respectively.
- $\text{Cor}(X, Y)$ measures the strength of the linear relationship between the X and Y data, with stronger relationships as $\text{Cor}(X, Y)$ heads towards -1 or 1 .
- $\text{Cor}(X, Y) = 0$ implies no linear relationship.