Topic 3

Correlation and Regression

Correlation

Outline

Analogies to Vector Algebra

Covariance

Correlation

Law of Cosines

Introduction

We now take a careful look at the nature of linear relationships found in the data used to construct a scatterplot.

- Correlation examines this relationship in a symmetric manner.
  - Correlation focuses primarily of association,
  - Consequently, correlation does not attempt to establish any cause and effect.
- Regression, considers the relationship of a response variable as determined by one or more explanatory variables.
  - Regression is a often used as a tool to establish causality.
  - Regression is designed to help make predictions.

Analogies to Vector Algebra

vectors	quantitative observations
$\mathbf{v} = (v_1, \ldots, v_n)$ $\mathbf{w} = (w_1, \ldots, w_n)$	$\mathbf{x} = (x_1, \ldots, x_n)$ $\mathbf{y} = (y_1, \ldots, y_n)$
norm-squared $\lvert\lvert\mathbf{v}\rvert\rvert^2 = \sum_{i=1}^n v_i^2$	variance $s_x^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$
norm $\lvert\lvert\mathbf{v}\rvert\rvert$	standard deviation s_x
inner product $\langle \mathbf{v}, \mathbf{w} \rangle = \sum_{i=1}^n v_i w_i$	covariance $\mathrm{cov}(\mathbf{x}, \mathbf{y}) = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$
cosine $\cos\theta = \frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\lvert\lvert\mathbf{v}\rvert\rvert\,\lvert\lvert\mathbf{w}\rvert\rvert}$	correlation $r = \frac{\mathrm{cov}(\mathbf{x},\mathbf{y})}{s_x s_y}$

Covariance

The covariance measures the linear relationship between a pair of quantitative measures $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ on the same sample of n individuals.

$$\mathrm{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

- A positive covariance means that the terms $(x_i - \bar{x})(y_i - \bar{y})$ in the sum are more likely to be positive than negative. This occurs whenever the x and y variables are more often both above or below the mean in tandem than not.
- A negative covariance means that the terms $(x_i - \bar{x})(y_i - \bar{y})$ in the sum are more likely to be negative than positive. This occurs when one of the variables is above its mean, the other is more often below.

The covariance of x with itself $\mathrm{cov}(x, x) = s_x^2$ is the variance of x.

Correlation

The correlation, r, is the covariance of the standardized versions of x and y.

$$\begin{aligned} r(x,y) &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) \\ &= \frac{\operatorname{cov}(x,y)}{s_x s_y}. \end{aligned}$$

The observations x and y are called uncorrelated if $r(x,y)=0$.

NB. Because the standardized score has no units, neither does the correlation.

Law of Cosines

By expanding the squares in the sum for the variance s_{x+y}, we have

$$\begin{aligned} s_{x+y}^2 &= \sum_{i=1}^{n}((x_i + y_i) - (\bar{x} - \bar{y}))^2 \\ &= s_x^2 + s_y^2 + 2\mathrm{cov}(x, y) \\ &= s_x^2 + s_y^2 + 2 r s_x s_y . \end{aligned}$$

Notice analogy between this formula and the law of cosines: $c^2 = a^2 + b^2 - 2ab\cos\theta$.

If the two observations are uncorrelated, we have the Pythagorean identity

$$s_{x+y}^2 = s_x^2 + s_y^2,$$

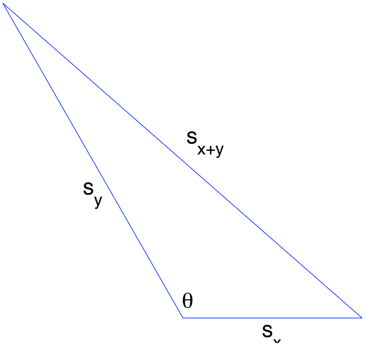

Figure: For the law of cosines, let $a = s_x$, $b = s_y$, $c = s_{x+y}$ and $r = -\cos\theta$

Correlation

We can use the law of cosines to see that correlation always takes on values between -1 and $+1$.

- First note that the correlation is the same for both the original observations and for the standardized observations, so we can assume that $s_x = s_y = 1$.
- Using the variance of $x + y$,

$$0 \leq s^2_{x+y} = s^2_x + s^2_y + 2rs_x s_y = 1 + 1 + 2r. \quad \text{So,} \quad -1 \leq r.$$

- Using the variance of $x - y$,

$$0 \leq s^2_{x-y} = s^2_x + s^2_y - 2rs_x s_y = 1 + 1 - 2r. \quad \text{So,} \quad r \leq +1.$$

- The values $r \pm 1$ occur precisely when x and y are perfectly linearly related. This is positively associated for $r = +1$ and negatively associated for $r = -1$.

Correlation

Example. *Archeopteryx*, which is generally accepted as being the oldest known bird, lived around 150 million years ago, in what is now southern Germany. The first complete specimen was announced in 1861, only two years after Charles Darwin published *On the Origin of Species*, and thus became a key piece of evidence in the debate over evolution. Below are the lengths in centimeters of the femur and humerus for the 5 specimens of *Archeopteryx* that have preserved both bones.

```
> femur<-c(38,56,59,64,74)
> humerus<-c(41,63,70,72,84)
> cor(femur,humerus)
[1] 0.9941486
```

This is very close to the maximum value of $+1$ and so the association is nearly perfectly positively linear.

Exercise. Make a scatterplot of the femur and humerus for the 5 specimens above. Compute by hand the correlation

Correlation

Exercise. Choose several value for `r` between -1 and $+1$ and look at the scatterplots to gain an intuitive feel for the value of the correlation.

For example, we can look at 100 observations with a correlation of $r = 0.5$ in R as follows:

```
> r<-0.5
> x<-rnorm(100)
> z<-rnorm(100)
> y<-r*x + sqrt(1-r^2)*z
> plot(x,y)
```

A 3×2 grid of plots can be obtained by entering `par(mfrow=c(3,2))` before the first plot.