Formulæ for concepts that Tom is drawing

Thomas Levine

March 22, 2014

These formulæ are for the bivariate (rather than more-than-two) populations (rather than samples).

1 Covariance

Conceptually, it's this.

$$\sigma_{xy} = \mathrm{E}\left[\,(x - \mathrm{E}[x])(y - \mathrm{E}[y])\,\right]$$

More precisely, it could be this.

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})\,(y_i - \bar{y})$$

That's the sum ($\sum_{i=1}^{N}$) of rectangles ($(x_i - \bar{x})\,(y_i - \bar{y})$) divided by the number of observations/rectangles ($\frac{1}{N}$).

As matrix arithmetic, it's written in a way that works for more than two variables. (Σ is the covariance matrix.)

$$\Sigma = \mathrm{E}\left[(\mathbf{X} - \mathrm{E}[\mathbf{X}])\,(\mathbf{X} - \mathrm{E}[\mathbf{X}])^{\mathrm{T}}\right]$$

It's multiplication of many $\mathbf{X}$ by many $\mathbf{X}$, and that's lots of rectangles for lots of covariances.

In R, it's this.

```
# Two variables
cov(x,y)

# More variables
var(iris[-5])
cov(iris[-5])
```

2 Variance

Variance is the covariance of something with itself.

$$\sigma_x^2 = \sigma_{xx} = \mathrm{E}\left[\,(x - \mathrm{E}[x])^2\,\right]$$

More precisely, like the covariance, it could be this.

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})$$

And that reduces to this.

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

This time, the top is a sum of squares ($(x_i - \bar{x})^2$).

As matrix arithmetic, it's the diagonal of the covariance because those are the cells where we're multiplying the the same $\mathbf{x}$.

$$diag\,(\Sigma)$$

In R, all of these would work.

```
# One variable
var(x)
cov(x,x)

# More variables
lapply(iris[-5],var)
diag(cov(iris[-5]))
```

3 Correlation

Correlation (specifically, the Pearson product-moment correlation coefficient—there are others) measures how much the variables linearlly depend on each other.

For perfectly linear variables that move together, ρ_{xy} will be 1. For perfectly linear variables that move oppositely, ρ_{xy} will be -1.

ρ_{xy} is just the covariance (σ_{xy}) divided by the product of the standard deviations ($\sigma_x\sigma_y$).

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

We can also factor out the N term and write it as the sum of the rectangles divided by the rectangle formed by the square roots of the sums of the squares.

$$\rho = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2}}$$

If you need to be convinced that σ_{xy} is no greater than $\sigma_x \sigma_y$, recall that the covariance includes negative rectangles and the variance does not;

4 Ordinary least-squares regression

Simple regression is a best fit line between two variables; we are looking for α and β.

$$\hat{y}_i = \alpha + \beta x_i$$

β is just the covariance divided by the variance.

$$\beta = \sigma xy / \sigma_x^2$$

We can also factor out the N term and write it as the sum of the rectangles divided by the sum of the squares.

$$\beta = \frac{\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2}$$

$$\beta = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

In matrix form, we might right this.

$$B = \left(X^T X\right)^{-1} X^T y$$

The part on the right $(X^T y)$ is multiplication of x-values by y-values, so it's the sum of rectangles like in covariance.

We invert the part on the left $\left(\left(X^T X\right)^{\mathbf{-1}}\right)$, so it's sort of like a denominator. We're multiplying x-values by x-values, so it's a sum of squares, like in variance.

5 Credits

Some formulæ were lifted from these pages.

- http://en.wikipedia.org/w/index.php?title=Covariance&action=edit
- http://en.wikipedia.org/w/index.php?title=Pearson_product-moment_correlation_coefficient&action=edit&section=3
- http://en.wikipedia.org/w/index.php?title=Ordinary_least_squares&action=edit