Review of Correlation

One of the most important tools in time series analysis is the Auto-Correlation Function, or ACF. In these gentle, introductory lectures that were developed for a review of basic statistics, we have discussed descriptive measures, both analytical and graphical, as well as statistical inference. In this lecture, we complete our overview of required concepts by looking at a very popular way to measure the strength of linear association between two variables, the *correlation coefficient*.

In your first probability or statistics course, you would have had an introduction to the ways in which we measure the strength of association of two variables. We'd guess that most people would first encounter the concept as the *Pearson product moment correlation coefficient*, often just called the correlation. Concepts like the Kullback–Leibler divergence probably come much later.

Let's start with the covariance concept. Recall that the variance of a single random variable is written, for the random variable X as

$$\sigma^2 \equiv V[X] \equiv E[\,(X-\mu_x)(X-\mu_X)\,]$$

For a data set we'd estimate this as

$$s^2 \equiv \frac{1}{n-1}\sum(x_i-\bar{x})(x_i-\bar{x})$$

This is really just an average, with the $n-1$ term to boost the estimate a little bit (for those who like unbiased estimators). Take a square root for the standard deviation to return to original units.

Now, if we have two random variables, we think about measuring their linear relationship with

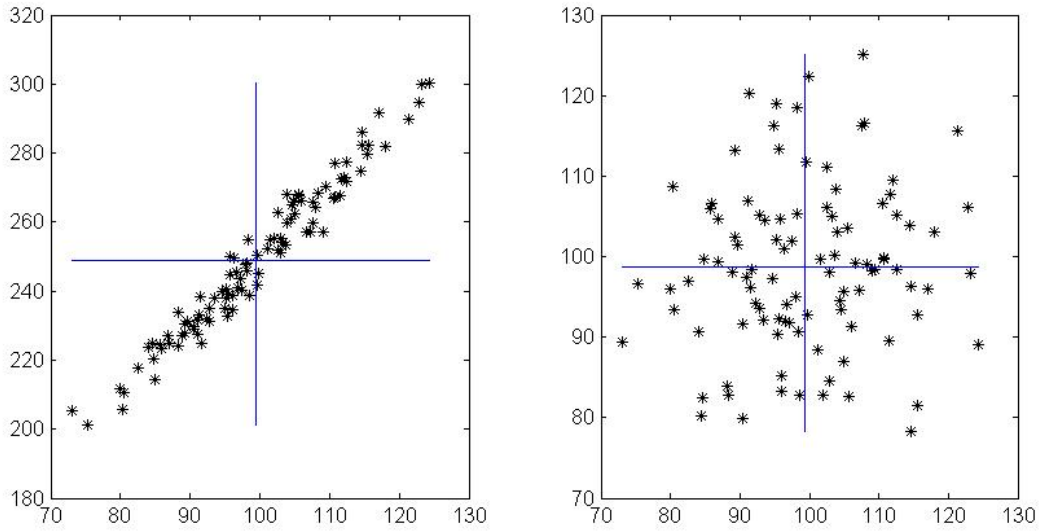
$$COV[X,Y] \equiv E[\,(X-\mu_X)(Y-\mu_Y)]$$

And, for data, we form the analogous estimator

$$cov \equiv \frac{1}{n-1}\sum(x_i-\bar{x})(y_i-\bar{y})$$

For motivation, we look at what happens "on average" (that's the expected value operator, $E[\quad]$) when we center the random variables, and then multiply these quantities together. Let's think through graphically why this is a good idea.

Most reasonable people would say that the y values given in the graph at the left track long with the x values quite well. That isn't true for the graph at the right. Another way of looking at this is to wonder whether a given y value can be decently well predicted by its corresponding x value.



If you look at the first graph you will see that most of the *above average* x values go along with the *above average* y values. It is the same thing with the below average x values and the below average y values. Think about the deviations (distance from the mean). This means that

- When $x_i - \bar{x} > 0$ it is pretty common to find $y_i - \bar{y} > 0$ as well. A positive times a positive is positive, so this means that $(x_i - \bar{x})(y_i - \bar{y}) > 0$ is also greater than 0.
- On the other side, when $x_i - \bar{x} < 0$ it is pretty common to find $y_i - \bar{y} < 0$ as well. A negative times a negative is negative, so this means that $(x_i - \bar{x})(y_i - \bar{y}) > 0$.
- There aren't many positive x values associated with negative y values. So there aren't many terms where $x_i - \bar{x} > 0$ and $y_i - \bar{y} < 0$. That means there aren't many terms where $(x_i - \bar{x})(y_i - \bar{y}) < 0$.
- Just to be complete, there aren't many negative x values associated with positive y values. So there aren't many terms where $x_i - \bar{x} < 0$ and $y_i - \bar{y} > 0$. Again, that means there aren't many terms where $(x_i - \bar{x})(y_i - \bar{y}) < 0$.

Summing up, in the first graph most of the products $(x_i - \bar{x})(y_i - \bar{y})$ give positive numbers and we would expect the sum of these terms to take us pretty far in the positive direction. We'd therefore expect $\sum(x_i - \bar{x})(y_i - \bar{y})$ to be on the large side.

When we move on to correlation, we are really just expressing the covariance concept in standard units. The motivation for this might be obvious- if we are measuring strength of linear association, we shouldn't have to worry about whether we've measure in feet, in inches, or in miles.

If we think about it in this way the defining formula for random variables should make sense:

$$\rho(X,Y) \equiv E\left[\left(\frac{X-\mu_X}{\sigma_x}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right]$$

If we have data instead of random variables, we estimate this term in the most direct way as

$$r \equiv \hat{\rho} \equiv \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_X}\right)\left(\frac{y_i-\bar{y}}{s_Y}\right)$$

Remember the "sum of squares" notation as follows:

$$SSX \equiv \sum(x_i-\bar{x})^2 = \sum x_i^2 - \frac{1}{n}\sum x_i \sum x_i$$

$$SSY \equiv \sum(y_i-\bar{y})^2 = \sum y_i^2 - \frac{1}{n}\sum y_i \sum y_i$$

$$SSXY \equiv \sum x_i\, y_i - \frac{1}{n}\sum x_i \sum y_i$$

We can rewrite more compactly using "sum of squares" notation:

$$\frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_X}\right)\left(\frac{y_i-\bar{y}}{s_Y}\right) = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{\sqrt{\frac{SSX}{n-1}}}\right)\left(\frac{y_i-\bar{y}}{\sqrt{\frac{SSY}{n-1}}}\right)$$

Cancel all those annoying $n-1$ terms, and pull constants through the sum:

$$r = \hat{\rho} = \sum\left(\frac{x_i-\bar{x}}{\sqrt{SSX}}\right)\left(\frac{y_i-\bar{y}}{\sqrt{SSY}}\right) = \frac{1}{\sqrt{SSX\ SSY}}\sum(x_i-\bar{x})(y_i-\bar{y}) = \frac{SSXY}{\sqrt{SSX}\sqrt{SSY}}$$

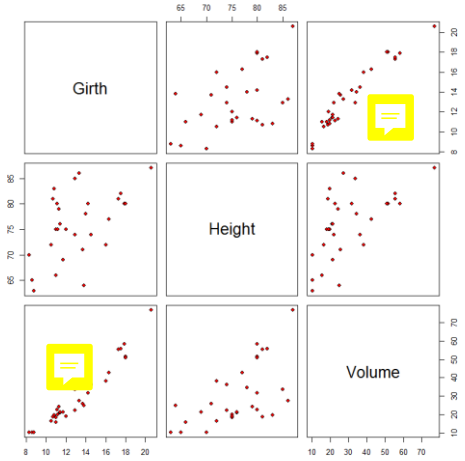
Girth, Height and Volume for Black Cherry Trees

R has a nice data set on some physical characteristics of Black Cherry Trees. While most of us think about girth as the distance *around* something, here we find a simple scaling being used.

> *This data set provides measurements of the girth, height and volume of timber in 31 felled black cherry trees. Note that girth is the diameter of the tree (in inches) measured at 4 ft 6 in above the ground.*

It's always natural to take a look at our data and we'll do so with a "pairs" plot.

```
pairs(trees, pch = 21, bg = c("red"))
```



It appears that our variables are all positively correlated, as we'd expect. The linear relationship between Girth and Volume seems especially strong. You wouldn't have come to that conclusion only looking at the covariances!

cov(trees)

	Girth	*Height*	*Volume*
Girth	*9.847914*	*10.38333*	*49.88812*
Height	*10.383333*	*40.60000*	*62.66000*
Volume	*49.888118*	*62.66000*	*270.20280*

cor(trees)

	Girth	*Height*	*Volume*
Girth	*1.0000000*	*0.5192801*	*0.9671194*
Height	*0.5192801*	*1.0000000*	*0.5982497*
Volume	*0.9671194*	*0.5982497*	*1.0000000*