Lecture 3

Autocorrelation
Textbook Sections: 1.3, 1.4

Covariance and Correlation Properties

When we're interested with the relationship of two random variables X and Y, we often start by looking at their covariance,

$$Cov(X,Y) = \mathrm{E}\left[(X - E(X))(Y - E(Y))\right],$$

and correlation,

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

Let $V_1, V_2, , ,$ and $W_1, W_2, ...$ be random variables, and $c_1, c_2, ...$and $d_1, d_2, ...$be constants.

(a) The following are true

$$\begin{aligned}
E(c_1V_1 + c_2V_2) &= c_1E(V_1) + c_2E(V_2), \\
Var(c_1V_1 + c_2V_2) &= c_1^2Var(V_1) + c_2^2Var(V_2) + 2c_1c_2Cov(V_1, V_2), \\
Cov(c_1V_1 + c_2V_2, d_1W_1 + d_2W_2) &= c_1d_1Cov(V_1, W_1) + c_1d_2Cov(V_1, W_2) \\
&\quad + c_2d_1Cov(V_2, W_1) + c_2d_2Cov(V_2, W_2).
\end{aligned}$$

(b) The following are generalizations of the results in (a)

$$\begin{aligned}
E(\sum c_iV_i) &= \sum c_iE(V_i), \\
Var(\sum c_iV_i) &= \sum c_i^2Var(V_i) + \sum_i \sum_{j \neq i} c_ic_jCov(V_i, V_j) \\
&= \sum_i \sum_j c_ic_jCov(V_i, V_j), \\
Cov(\sum c_iV_i, \sum d_jW_j) &= \sum_i \sum_j c_id_jCov(V_i, W_j).
\end{aligned}$$

(c) A consequence of the results in part (b) is that if V_i are mutually uncorrelated (i.e., $Cov(V_i, V_j) = 0$ whenever $i \neq j$), then

$$Var(\sum c_iV_i) = \sum c_i^2Var(V_i).$$

It is important to keep in mind that, for any random variable V, $Cov(V, V) = Var(V)$.

Notation

$Cov(X,Y)$ is denoted by $\sigma(X,Y)$ or σ_{XY}.
$Corr(X,Y)$ is denoted by $\rho(X,Y)$ or ρ_{XY}. This is the Greek letter rho (pronounced like "row").

Sample Estimates

Estimates of variance and covariance are:

$$\widehat{Cov(X,Y)} = S_{\text{XY}}/(n-1), \qquad \widehat{Var(X)} = S_{\text{XX}}/(n-1), \qquad \widehat{Var(Y)} = S_{\text{YY}}/(n-1),$$

where

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

$$S_{\text{XX}} = \sum_{i=1}^{n}(X_i - \bar{X})^2, \qquad S_{\text{YY}} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2, \qquad S_{\text{XY}} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}).$$

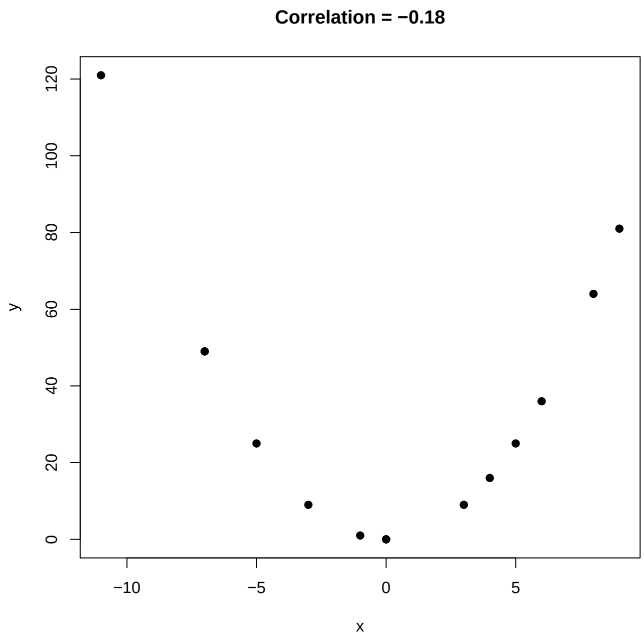
Plugging in these estimates we get the sample correlation coefficient:

$$\hat{\rho} = \frac{S_{XY}/(n-1)}{\sqrt{[S_{XX}/(n-1)][S_{YY}/(n-1)]}} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}.$$

In R, given two vectors `x` and `y`, the sample covariance and correlation can be obtained by `cov(x,y)` and `cor(x,y)`, respectively.

Understanding Correlation

Correlation measures the strength of the *linear* relationship between X and Y.
A low correlation coefficient indicates that there is a weak linear relationship, but there may be a nonlinear relationship between X and Y.



Autocovariance and Autocorrelation

In time series analysis, we deal with a sequence of variables $\{X_t\}$. We want to understand how the variables are related so that we can model the structure and make forecasts. In order to do this, we would like to know $Cov(X_t, X_s)$ and $Corr(X_t, X_s)$ for different indices t and s.

However, if we consider all of these different covariances or correlations, we will have too many parameters to estimate. We need a simplified framework.

We restrict ourselves to processes, for which the correlation depends not on the exact indices s and t, but only on the difference between the indices $s - t$. This difference is called the lag. Instead of looking at $Corr(X_t, X_s)$ for all possible pairs of t and s, we only need to look at $Corr(X_t, X_{t+1}), Corr(X_t, X_{t+2}), Corr(X_t, X_{t+3})$, etc.

Simplifying the notation, we have the autocovariance function,

$$\gamma(h) = Cov(X_t, X_{t+h}),$$

and the autocorrelation function,

$$\rho(h) = Corr(X_t, X_{t+h}) = \gamma(h)/\gamma(0).$$

So $\gamma(1)$ is the covariance of random variables that are one time point apart,
$\gamma(2)$ is the covariance of random variables that are two time points apart, and so on.

Keep in mind that $\gamma(0) = Cov(X_t, X_t) = Var(X_t)$ and $\rho(0) = 1$.

Here is another important fact:

$$\gamma(-h) = Cov(X_{t+h}, X_t) = Cov(X_t, X_{t+h}) = \gamma(h),\ h = 0, 1, ...,$$
$$\rho(-h) = Corr(X_{t+h}, X_t) = Corr(X_t, X_{t+h}) = \rho(h),\ h = 0, 1, ...$$

This fact tells us it is enough to investigate the autocorrelation $\rho(h)$ for nonnegative integer h and there is no need to look at $\rho(-h)$.

Examples

1. If $\{X_t\}$ i.i.d. (or at least uncorrelated) with mean μ and variance σ^2, then

$$\gamma(h) = Cov(X_t, X_{t+h}) = \begin{cases} 0 & h = 1, 2, ... \\ \sigma^2 & h = 0, \end{cases}$$

$$\rho(h) = Corr(X_t, X_{t+h}) = \begin{cases} 0 & h = 1, 2, ... \\ 1 & h = 0. \end{cases}$$

2. If a sequence $\{X_t\}$ has the structure

$$X_t = \varepsilon_t + \theta\varepsilon_{t-1},$$

where the sequence $\{\varepsilon_t\}$ are i.i.d. with mean zero and variance σ^2, then $\{X_t\}$ is called a **moving average of order 1**, or simply **MA(1)**. For all t we have:

$$E(X_t) = 0,$$

$$Var(X_t) = (1 + \theta^2)\sigma^2,$$

$$\gamma(h) = Cov(X_t, X_{t+h}) = \begin{cases} (1 + \theta^2)\sigma^2 & h = 0 \\ \theta\sigma^2 & h = 1 \\ 0 & h \geq 2. \end{cases}$$

$$\rho(h) = \begin{cases} 1 & h = 0 \\ \theta/(1 + \theta^2) & h = 1 \\ 0 & h \geq 2. \end{cases}$$

Since $\{\varepsilon_i\}$ are mutually uncorrelated, we have

$$\begin{aligned} \gamma(0) &= Var(X_t) = Var(\varepsilon_t + \theta\varepsilon_{t-1}) = Var(\varepsilon_t) + \theta^2 Var(\varepsilon_{t-1}) \\ &= \sigma^2 + \theta^2\sigma^2 = (1 + \theta^2)\sigma^2, \\ \gamma(1) &= Cov(X_t, X_{t+1}) = Cov(\varepsilon_t + \theta\varepsilon_{t-1}, \varepsilon_{t+1} + \theta\varepsilon_t) \\ &= Cov(\varepsilon_t, \varepsilon_{t+1}) + \theta Cov(\varepsilon_t, \varepsilon_t) + \theta Cov(\varepsilon_{t-1}, \varepsilon_{t+1}) + \theta^2 Cov(\varepsilon_{t-1}, \varepsilon_t) \\ &= 0 + \theta\sigma^2 + 0 + 0 = \theta\sigma^2, \\ \gamma(2) &= Cov(X_t, X_{t+2}) = Cov(\varepsilon_t + \theta\varepsilon_{t-1}, \varepsilon_{t+2} + \theta\varepsilon_{t+1}) = 0. \end{aligned}$$

Note that $\gamma(2) = 0$ since there is no common ε-term in X_t and X_{t+2}. Incidentally, the same argument applies for $\gamma(3), \gamma(4), ...$ Since $\rho(h) = \gamma(h)/\gamma(0)$, we have

$$\begin{aligned} &\rho(0) = 1, \\ &\rho(1) = [\theta\sigma^2]/[(1 + \theta^2)\sigma^2] = \theta/(1 + \theta^2), \\ &\rho(2) = 0/\gamma(0) = 0, \rho(3) = 0, ... \end{aligned}$$

3. A sequence is called a **random walk** if it has the form $X_t = X_{t-1} + \varepsilon_t$, where $\{\varepsilon_t\}$ are i.i.d. with mean zero and variance σ^2. For this sequence

$$Cov(X_t, X_{t+h}) = t\sigma^2, \; h \geq 1.$$