STAT 520: Forecasting and Time Series

David B. Hitchcock
University of South Carolina
Department of Statistics

What are Time Series Data?

- Time series data are collected sequentially over time.
- Some common examples include:
  1. Meteorological data (temperatures, precipitation levels, etc.) taken daily or hourly
  2. Sales data taken annually
  3. Heart activity measured at each millisecond
- The major goals of time series analysis are: (1) to model the stochastic phenomenon that produces these data; and (2) to predict or forecast the future values of the time series.
- A key aspect of time series data is that observations measured across time are typically *dependent random variables*, not independent r.v.'s.

Some Time Series Examples

- See annual Los Angeles rainfall plot in R. There is substantial variation in rainfall amount across years.
- Are consecutive years related? Can we predict a year's rainfall amount based on the previous year's amount?
- Scatter plot of one year's rainfall against previous year's rainfall shows little association.
- See color property value plot in R. Can we predict a batch's color property based on the previous batch's value?
- Scatter plot of one batch's color value against previous batch's value shows some positive association.
- See Canadian hare abundance plot in R. Can we predict a year's abundance based on the previous year's abundance?
- Scatter plot of one year's abundance against previous year's abundance shows clear positive association.

Some Time Series Examples with Seasonality

- See monthly Dubuque temperature plot in R. Notice the regular pattern.
- These data show *seasonality*: we see that observations that are 12 months apart are related.
- In the Dubuque data, each January temperature is low, while each July temperature is high.
- A seasonal pattern is also seen in the monthly oil filter sales data.
- Certain months regularly see high sales while other months regularly see low sales.

Some Examples with Multiple Time Series

- See Southern Oscillation Index (SOI) and Recruitment time series plots.
- We may investigate how SOI and the fish population are related over time.
- The fMRI time series plots show several similar time series taken under different experimental conditions.

- An observed time series can be modeled with a *stochastic process*: a sequence of random variables taken across time $\{Y_t, t = \ldots, -2, -1, 0, 1, 2, \ldots\}$.
- The probability structure of a stochastic process is determined by the set of all joint distributions of all finite sets of these r.v.'s.
- If these joint distributions are *multivariate normal*, it is simpler: Knowing the means, variances, and covariances of the Y_t's tells us everything about the joint distributions.

Moments (Means, Variances, Covariances)

- The mean function of a stochastic process $\{Y_t\}$

$$\mu_t = E(Y_t) \text{ for all } t$$

  gives the expected value of the process at time t.
- μ_t could vary across time.
- The autocovariance function is denoted:

$$\gamma_{t,s} = cov(Y_t, Y_s) \text{ for } t, s$$

  where
  $cov(Y_t, Y_s) = E[(Y_t - \mu_t)(Y_s - \mu_s)] = E(Y_t Y_s) - \mu_t \mu_s$.
- The autocorrelation function is denoted:

$$\rho_{t,s} = corr(Y_t, Y_s) \text{ for } t, s$$

  where $corr(Y_t, Y_s) = cov(Y_t, Y_s)/[(var(Y_t)var(Y_s))^{1/2}]$.

Interpreting Correlations and Covariances

- Both autocovariance and autocorrelation measure the linear dependence between the process's values at two different times.
- The autocorrelation is scaled to be between -1 and 1 and is easier to interpret.
- Many time series processes have positive autocovariance and autocorrelation:
  - If $\gamma_{t,s} > 0$, then: If Y_t is large (small), then Y_s tends to be large (small).
  - If $\gamma_{t,s} < 0$, then: If Y_t is large, then Y_s tends to be small (and vice versa).
- Note $\gamma_{t,t} = var(Y_t) = E[(Y_t - \mu_t)^2] = E(Y_t^2) - \mu_t^2$.

Important Results

$$cov\left[\sum_{i=1}^{m} c_i Y_{t_i}, \sum_{j=1}^{n} d_j Y_{s_j}\right] = \sum_{i=1}^{m}\sum_{j=1}^{n} c_i d_j cov(Y_{t_i}, Y_{s_j}).$$

A special case of this result:

$$var\left[\sum_{i=1}^{n} c_i Y_{t_i}\right] = \sum_{i=1}^{n} c_i^2 var(Y_{t_i}) + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} c_i c_j cov(Y_{t_i}, Y_{t_j}).$$

Simple Example of Calculating a Correlation of Linear Combination of r.v.'s

- **Example:** Let Y_1 be a r.v. with $E(Y_1) = 0$ and $var(Y_1) = 4$ and Y_2 be a r.v. with $E(Y_2) = 0$ and $var(Y_2) = 9$, and let $cov(Y_1, Y_2) = 1$.
- Find $corr(2Y_1 + Y_2, 3Y_1 - 2Y_2)$.
- To calculate the correlation between two random quantities: First calculate the covariance between the two quantities; then calculate the variance of each quantity; and then plug into the formula for correlation.

Continuation of of Calculating a Correlation

- **Covariance:**
- $cov(2Y_1 + Y_2, 3Y_1 - 2Y_2) =$
  $6cov(Y_1, Y_1) - 4cov(Y_1, Y_2) + 3cov(Y_2, Y_1) - 2cov(Y_2, Y_2) =$
  $6(4) - 4(1) + 3(1) - 2(9) = 24 - 4 + 3 - 18 = 5.$
- **Variances:**
- $var(2Y_1 + Y_2) = cov(2Y_1 + Y_2, 2Y_1 + Y_2) =$
  $4var(Y_1) + 2cov(Y_1, Y_2) + 2cov(Y_2, Y_1) + var(Y_2) =$
  $4(4) + 2(1) + 2(1) + 9 = 29.$
- $var(3Y_1 - 2Y_2) = cov(3Y_1 - 2Y_2, 3Y_1 - 2Y_2) =$
  $9var(Y_1) - 6cov(Y_1, Y_2) - 6cov(Y_2, Y_1) + 4var(Y_2) =$
  $9(4) - 6(1) - 6(1) + 4(9) = 60.$
- **Formula for Correlation:**
- $corr(2Y_1 + Y_2, 3Y_1 - 2Y_2) = 5/\sqrt{(29)(60)} \approx 0.12.$

A simple time series process: The Random Walk

- Let $e_1, e_2, \ldots$ be a sequence of independent and identically distributed (iid) r.v.'s, each having mean 0 and variance σ_e^2.
- Consider the time series:

$$\begin{aligned} Y_1 &= e_1 \\ Y_2 &= e_1 + e_2 \\ &\vdots \\ Y_t &= e_1 + e_2 + \cdots + e_t \end{aligned}$$

- In other words, $Y_t = Y_{t-1} + e_t$, where initially $Y_1 = e_1$.
- Then Y_t is the position on a number line (at time t) of a walker who is taking random (forward or backward) steps (of sizes $e_1, e_2, \ldots, e_t$) along the number line.

Properties of the Random Walk

- For this random walk process, the mean function $\mu_t = 0$ for all t, since

$$E(Y_t) = E(e_1 + e_2 + \cdots + e_t) = 0 + 0 + \cdots + 0 = 0$$

and

$$var(Y_t) = var(e_1 + e_2 + \cdots + e_t) = \sigma_e^2 + \sigma_e^2 + \cdots + \sigma_e^2 = t\sigma_e^2$$

since all the e_i's are independent.

Autocovariance Function of the Random Walk

- For $1 \leq t \leq s$,

$$\begin{aligned}\gamma_{t,s} &= cov(Y_t, Y_s) \\ &= cov(e_1 + e_2 + \cdots + e_t, \\ &\quad e_1 + e_2 + \cdots + e_t + e_{t+1} + \cdots + e_s)\end{aligned}$$

- From the formula for covariance of sums of r.v.'s, we have

$$cov(Y_t, Y_s) = \sum_{i=1}^{s} \sum_{j=1}^{t} cov(e_i, e_j),$$

but these covariance terms are zero except when $i = j$, so $cov(Y_t, Y_s) = t\sigma_e^2$, for $1 \leq t \leq s$.

Autocorrelation Function of the Random Walk

- The autocorrelation function is easily found to be $\sqrt{t/s}$, for $1 \leq t \leq s$.
- From this, note $corr(Y_1, Y_2) = \sqrt{1/2} = 0.707$;
  $corr(Y_{24}, Y_{25}) = \sqrt{24/25} = 0.98$;
  $corr(Y_1, Y_{25}) = \sqrt{1/25} = 0.20$.
- Values close in time are more strongly correlated than values far apart in time.
- And neighboring values later in the process are more strongly correlated than neighboring values early in the process.

A Simple Moving Average Process

- Let $e_1, e_2, \ldots$ be a sequence of independent and identically distributed (iid) r.v.'s, each having mean 0 and variance σ_e^2.
- Consider the time series:

$$Y_t = \frac{e_t + e_{t-1}}{2}.$$

- For this moving average process, the mean function $\mu_t = 0$ for all t, since $E(Y_t) = E[(e_t + e_{t-1})/2] = 0.5E[e_t + e_{t-1}] = 0$, and $var(Y_t) = var[(e_t + e_{t-1})/2] = 0.25 var[e_t + e_{t-1}] = 0.25 \times 2\sigma_e^2 = 0.5\sigma_e^2$ since all the e_i's are independent.

Autocovariance Function of the Moving Average

$$\begin{aligned} cov(Y_t, Y_{t-1}) &= cov\left(\frac{e_t + e_{t-1}}{2}, \frac{e_{t-1} + e_{t-2}}{2}\right) \\ &= 0.25[cov(e_t, e_{t-1}) + cov(e_t, e_{t-2}) + \\ &\quad cov(e_{t-1}, e_{t-1}) + cov(e_{t-1}, e_{t-2})] \\ &= 0.25[0 + 0 + cov(e_{t-1}, e_{t-1}) + 0] = 0.25\sigma_e^2 \end{aligned}$$

- And

$$cov(Y_t, Y_{t-2}) = cov\left(\frac{e_t + e_{t-1}}{2}, \frac{e_{t-2} + e_{t-3}}{2}\right) = 0,$$

  since there are no overlapping e terms here, and all the e_i's are independent.
- Similarly, $cov(Y_t, Y_{t-k}) = 0$ for all $k > 1$.

Autocorrelation Function of the Moving Average

- From this, note

$$\rho_{t,s} = \begin{cases} 1, \text{for } |t-s| = 0 \\ 0.5, \text{for } |t-s| = 1 \\ 0, \text{for } |t-s| > 1 \end{cases}$$

- So $\rho_{t,t-1}$ is the same no matter what t is, and in fact, for any k, $\rho_{t,t-k}$ is the same no matter what t is.
- This is related to the concept of *stationarity*.

Stationarity

- If a process is stationary, this implies that the laws that govern the process do not change as time goes on.
- A process is *strictly stationary* if the entire joint distribution of n values is the same as the joint distribution of any other n time-shifted values of the process, no matter when the two sequences start.
- For example, with a stationary process, the joint distribution of Y_1, Y_3, Y_4 would be the same as the joint distribution of Y_6, Y_8, Y_9, and similarly for any such pairs of sequences.

Properties of Stationarity Processes

- Note that with a stationary process: $E(Y_t) = E(Y_{t-k})$ for all t and k, so this implies that the mean function of any stationary process is *constant* over time.
- Also, with a stationary process: $var(Y_t) = var(Y_{t-k})$ for all t and k, so the variance function of any stationary process is *constant* over time.
- Note: A function that is constant over time is one that does not depend on t.

More on Stationarity

- Also, if the process is stationary, the bivariate distribution of (Y_t, Y_s) is the same as the bivariate distribution of (Y_{t-k}, Y_{s-k}) for all t, s, k.
- So $cov(Y_t, Y_s) = cov(Y_{t-k}, Y_{s-k})$ for all t, s, k.
- Letting $k = s$, we have $cov(Y_t, Y_s) = cov(Y_{t-s}, Y_0)$; letting $k = t$, we have $cov(Y_t, Y_s) = cov(Y_0, Y_{s-t})$.
- So $cov(Y_t, Y_s) = cov(Y_0, Y_{|t-s|})$.
- So for a stationary process, the covariance between any two values depends only on the *lag* in time between the values, not on the actual times t and s.
- For a stationary process, we can express our autocovariance and autocorrelation functions simply in terms of the time lag k:
  $\gamma_k = cov(Y_t, Y_{t-k})$ and $\rho_k = corr(Y_t, Y_{t-k})$.

Weak Stationarity

- A process is *weakly stationary* or *second-order stationary* if
  1. The mean function is constant over time, and
  2. $\gamma_{t,t-k} = \gamma_{0,k}$ for every time t and lag k
- Any process that is strictly stationary is also weakly stationary.
- But a process could be weakly stationary and NOT strictly stationary.
- In the special case that all joint distributions for the process are multivariate normal, then being weakly stationary is *equivalent* to being strictly stationary.

White Noise

- The *white noise process* is a simple example of a stationary process.
- White noise is simply a sequence of iid r.v.'s $\{e_t\}$.
- White noise is strictly stationary since

$$\begin{aligned} & P[e_{t_1} \leq x_1, \ldots, e_{t_n} \leq x_n] \\ = \; & P[e_{t_1} \leq x_1] \cdots P[e_{t_n} \leq x_n] \\ = \; & P[e_{t_1 - k} \leq x_1] \cdots P[e_{t_n - k} \leq x_n] \\ = \; & P[e_{t_1 - k} \leq x_1, \ldots, e_{t_n - k} \leq x_n] \end{aligned}$$

- Clearly, $\mu_t = E(e_t)$ is constant, and $\gamma_k = var(e_t) = \sigma_e^2$ for $k = 0$ and zero for any $k \neq 0$.

Examples of Stationary and Nonstationary Processes

- The moving average process is another stationary process.
- The random walk process is not stationary. How can we see that?
- Its variance function is NOT constant, and its autocovariance function does NOT only depend on the time lag.
- What if we considered the *differences* of successive Y-values: $\nabla Y_t = Y_t - Y_{t-1}$?
- Since for the random walk, $\nabla Y_t = e_t$, or simply white noise, we see the *differenced* time series is stationary.
- This is common in practice: We can often transform nonstationary processes into stationary processes by *differencing*.