Time Series Analysis and its Application

An Outline of "TSA and its Applications" by Shumway and Stoffer

Darshan Patel

August 18, 2019

Contents

1 Characteristic of Time Series

1.1 The Nature of Time Series Data

- There are two different approaches for time series analysis: the time domain approach and the frequency domain approach
- The time domain approach views the investigation of lagged relationships as most important (e.g., how does what happened today affect what will happen tomorrow)
- The frequency domain approach views the investigation of cycles as most important (e.g., what is the economic cycle through periods of expansion and recession)
- Common cases of the experimental time series data include: quarterly earnings, global warming (temperature), speech data, financial data, population, imaging, natural occurrences, and more

1.2 Time Series Statistical Models

- Assume that a time series can be defined as a collection of random variables indexed according to the order they are obtained in time
- A collection of random variables $\{x_t\}$, indexed by t, is referred to as a stochastic process
- The observed values of a stochastic process are referred to as a realization of the stochastic process
- It is conventional to display a sample time series graphically by plotting the values of the random variables on the vertical axis, or ordinate, with the time scale as the abscissa; it is usually convenient to connect the values at adjacent time periods to reconstruct visually some original hypothetical continuous time series that might have produced these values as a discrete sample
- Continuous time series refer to series that could have been observed at any continuous point; the approximation of these series by discrete time parameter series sampled at equally spaced points in time acknowledges that the sampled data will be discrete because of restrictions inherent in the method of collection
- Time series can be distinguished by smoothness; smoothness is induced by the supposition that adjacent points in time are correlated, so the value of the series at time t, x_t, depends in some way on the past values $x_{t-1}, x_{t-2}, \ldots$
- A simple kind of generated series is a collection of uncorrelated random variables w_t with mean 0 and finite variance σ_w^2; the time series generated from uncorrelated variables is commonly used as a model for noise and is called white independent noise
- It is sometimes required that the noise is independent and identically distributed (iid) random variables with mean 0 and variance σ_w^2; this is distinguished by writing $w_t \sim \text{iid}(0, \sigma_w^2)$ or by saying white independent noise or iid noise

- A particular useful white nose series is Gaussian white noise where the w_t are independent normal random variables with mean 0 and variance σ_w^2, or more succinctly, $w_t \sim N(0, \sigma_w^2)$
- If the stochastic behavior of all time series could be expressed in terms of the white noise model, classical statistical methods would suffice
- Two ways of introducing serial correlation and more smoothness into time series models is moving averages and autoregression
- Moving averages uses the idea of replacing w_t with an average of its current value and its immediate neighbors in the past and future

  $$v_t = \frac{1}{3}(w_{t-1} + w_t + w_{t+1})$$

  A linear combination of values in a time series as above is referred to, generically, as a filtered series
- An autoregression model represents a regression of prediction of the current value x_t of a time series as a function of the past two values (or more) of the series, such as the following

  $$x_t = x_{t-1} - .9x_{t-2} + w_t$$

  successively for $t = 1, 2, \ldots$; a problem with startup values exists here because it also depends on the initial conditions x_0 and x_{-1}
- A model for analyzing trends such as seen in global temperature is the random walk with drift model given by

  $$x_t = \delta + x_{t-1} + w_t$$

  for $t = 1, 2, \ldots$ with initial condition $x_0 = 0$ and where w_t is white noise; the constant δ is called the drift and when $\delta = 0$, it is simply a random walk
- The term random walk comes from the fact that when $\delta = 0$, the value of the time series at time t is the value of the series at time $t - 1$ plus a completely random movement determined by w_t; therefore the above equation for random walk with drift can be written as

  $$x_t = \delta t + \sum_{j=1}^{t} w_j$$

  for $t = 1, 2, \ldots$
- Many realistic models for generating time series assume an underlying signal with some consistent periodic variation, contaminated by adding a random noise; the ratio of the amplitude of the signal to σ_w is sometimes called the signal to noise ratio (SNR); the larger the SNR, the easier it is to detect the signal

- Spectral analysis is a technique for detecting regular or periodic signals; it emphasizes the importance of simple additive models in the form of

$$x_t = s_t + v_t$$

where s_t denotes some unknown signal and v_t denotes a time series that may be white or correlated over time

1.3 Measures of Dependence

- A complete description of a time series, observed as a collection of n random variables at arbitrary time points $t_1, t_2, \ldots, t_n$, for any positive integer n, is provided by the joint distribution function, evaluated as the probability that the values of the series are jointly less than the n constants $c_1, c_2, \ldots, c_n$

$$F_{t_1,t_2,\ldots,t_n}(c_1, c_2, \ldots, c_n) = P(x_{t_1} \leq c_1, x_{t_2} \leq c_2, \ldots, x_{t_n} \leq c_n)$$

- Although the joint distribution function describes the data completely, it must be evaluated as a function of n arguments, so any plotting of the corresponding multivariate density functions is impossible

- The marginal distribution functions

$$F_t(x) = P(x_t \leq x)$$

of the corresponding marginal density functions

$$f_t(x) = \frac{\partial F_t(x)}{\partial x}$$

when they exist, are often informative for examining the marginal behavior of a series

- If x_t is Gaussian with mean μ_t and variance σ_t^2, abbreviated as $x_t \sim N(\mu_t, \sigma_t^2)$, the marginal density is given by

$$f_t(x) = \frac{1}{\sigma_t\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma_t^2}(x - \mu_t)^2\right\}$$

where $x \in \mathbb{R}$

- The mean function is defined as

$$\mu_{xt} = \mathrm{E}[x_t] = \int_{-\infty}^{\infty} x f_t(x)\, dx$$

provided it exists, where E denotes the expected value operator

- If w_t denotes a white noise series, then $\mu_{wt} = \mathrm{E}[w_t] = 0$ for all t; smoothing the series does not change the mean

- For a random walk with drift model, because $\mathrm{E}[w_t] = 0$ for all f and δ is a constant,

$$\mu_{xt} = \mathrm{E}[x_t] = \delta t + \sum_{j=1}^{t} \mathrm{E}[w_j] = \delta t$$

  which is a straight line with slope δ

- The autocovariance function is defined as the second moment product

$$\gamma_x(s,t) = \mathrm{Cov}[x_s,\ x_t] = \mathrm{E}[(x_s - \mu_s)(x_t - \mu_t)]$$

  for all s and t; when no possible confusion exists about which time series being referred to, drop the subscripts and write $\gamma_x(s,t)$ as $\gamma(s,t)$; note that $\gamma_x(s,t) = \gamma_x(t,s)$ for all time points s and t

- The autocovariance measures the linear dependence between two points on the same series observed at different times; very smooth series exhibit autocovariance functions that stay large even when the t and s are far apart, whereas choppy series tend to have autocovariance functions that are nearly zero for large separations

- If $\gamma_x(s,t) = 0$ and x_s and x_t are not linearly related, there may still be some dependence structure between them; if however x_s and x_t are bivariate normal, $\gamma_x(s,t) = 0$ ensures their independence

- It is clear that, for $s = t$, the autocovariance reduces to the variance, because

$$\gamma_x(t,t) = \mathrm{E}\left[(x_t - \mu_t)^2\right] = \mathrm{Var}[x_t]$$

- Covariance of Linear Combinations: if the random variables

$$U = \sum_{j=1}^{m} a_j X_j \text{ and } V = \sum_{k=1}^{r} b_k Y_k$$

  are linear combinations of (finite variance) random variables $\{X_j\}$ and $\{Y_k\}$, respectively, then

$$\mathrm{Cov}[U,\ V] = \sum_{j=1}^{m}\sum_{k=1}^{r} a_j b_k \mathrm{Cov}[X_j,\ Y_k]$$

  Furthermore, $\mathrm{Var}[U] = \mathrm{Cov}[U,\ U]$

- The smoothing operation introduces a covariance function that decreases as the separation between the two time points increases and disappears completely when the time points are separated by three or more time points

- For the random walk model, $x_t = \sum_{j=1}^{t} w_j$,

$$\gamma_x(s,t) = \mathrm{Cov}[x_s,\ x_t] = \mathrm{Cov}\left[\sum_{j=1}^{s} w_j,\ \sum_{k=1}^{t} w_k\right] = \min\{s,t\}\sigma_w^2$$

because the w_t are uncorrelated random variables; the autocovariance function of a random walk depends on the particular time values s and t and not on the time separation or lag; also note that the variance of the random walk, $\text{Var}[x_t] = \gamma_x(t,t) = t\sigma_w^2$, increases without bound as time t increases

- The autocorrelation function (ACF) is defined as

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}}$$

The ACF measures the linear predictability of the series at time t using only the value x_s

- It can be shown that $-1 \leq \rho(s,t) \leq 1$; if x_t is predicted perfectly from x_s through a linear relationship, $x_t = \beta_0 + \beta_1 x_s$, then the correlation would be $+1$ when $\beta_1 > 0$ and -1 when $\beta_1 < 0$

- The cross-covariance function between two series x_t and y_t is

$$\gamma_{xy}(s,t) = \text{Cov}[x_s,\ y_t] = \text{E}[(x_s - \mu_{xs})(y_t - \mu_{yt})]$$

- The cross-correlation function (CCF) is given by

$$\rho_{xy}(s,t) = \frac{\gamma_{xy}(s,t)}{\sqrt{\gamma_x(s,s)\gamma_y(t,t)}}$$

- The above ideas can be extended to the case of more than two series, $x_{t1}, x_{t2}, \ldots, x_{tr}$, or multivariate time series with r components

1.4 Stationary Time Series

- A strictly stationary time series is one for which the probabilistic behavior of every collection of values

$$\{x_{t_1}, x_{t_2}, \ldots, x_{t_k}\}$$

is identical to that of the time shifted set

$$\{x_{t_1+h}, x_{t_2+h}, \ldots, x_{t_k+h}\}$$

That is,

$$P(x_{t_1} \leq c_1, \ldots, x_{t_k} \leq c_k) = P(x_{t_1+h} \leq c_1, \ldots, x_{t_k+h} \leq c_k)$$

for all $k = 1, 2, \ldots$, all time points $t_1, t_2, \ldots, t_k$, all numbers $c_1, c_2, \ldots, c_k$ and all time shifts $h = 0, \pm 1, \pm 2, \ldots$

- If a time series is strictly stationary, then all of the multivariate distribution functions for subsets of variables must agree with their counterparts in the shifted set for all values of the shift parameter h

- A weakly stationary time series x_t is a finite variance process such that

  1. the mean value function μ_t is constant and does not depend on time t
  2. the autocovariance function $\gamma(s,t)$ depends on s and t only through their difference $|s-t|$

  Hence the term stationary will be used to mean weakly stationary; if a process is stationary in the strict sense, the term strictly stationary will be used

- Stationarity requires regularity in the mean and autocorrelation functions that these quantities may be estimated by averaging

- Because the mean function $\mathrm{E}[x_t] = \mu_t$ of a stationary time series is independent of time t, write $\mu_t = \mu$; also because the autocovariance function $\gamma(s,t)$ of a stationary time series x_t depends only on the difference $|s-t|$, then it can be rewritten as

$$\gamma(t+h,t) = \mathrm{Cov}[x_{t+h},\ x_t] = \mathrm{Cov}[x_h,\ x_0] = \gamma(h,0)$$

  where h represents the time shift or lag between time s and time t; thus, the autocovariance function of a stationary time series does not depend on the time argument t

- The autocovariance function of a stationary time series is

$$\gamma(h) = \mathrm{Cov}[x_{t+h},\ x_t] = \mathrm{E}[(x_{t+h}-\mu)(x_t-\mu)]$$

- The autocorrelation function (ACF) of a stationary time series is

$$\rho(h) = \frac{\gamma(t+h,t)}{\sqrt{\gamma(t+h,t+h)\gamma(t,t)}} = \frac{\gamma(h)}{\gamma(0)}$$

- The mean and autocovariance functions of the white noise series are evaluated as $\mu_{wt} = 0$ and

$$\gamma_w(h) = \mathrm{Cov}[w_{t+h},\ w_t] = \begin{cases} \sigma_w^2 & \text{if } h = 0 \\ 0 & \text{if } h \neq 0 \end{cases}$$

  Thus, white noise is weakly stationary or stationary; if the white noise variates are also normally distributed or Gaussian, the series is strictly stationary; the autocorrelation function is given by

$$\rho_w(h) = \begin{cases} 1 & \text{if } h = 0 \\ 0 & \text{if } h \neq 0 \end{cases}$$

- A moving average process is stationary because the mean function $\mu_{vt} = 0$ and the autocovariance function is independent of time t; the ACF is symmetric about lag zero

- A random walk is not stationary because its autocovariance function, $\gamma(s,t) = \min\{s,t\}\gamma_w^2$, depends on time; in addition, the random walk with drift violates the condition that the mean function $\mu_{xt} = \gamma t$ is a not a function of time t

- A model can be considered as having stationary behavior around a linear trend when the mean function is not independent of time but the autocovariance function is independent of time; this behavior is sometimes called trend stationary
- The autocovariance function of a stationary process, $\gamma(h)$, is non-negative definite, ensuring that variances of linear combinations of the variates x_t will never be negative; that is, for any $n \geq 1$, and constants $a_1, \ldots, a_n$,

$$0 \leq \text{Var}\left[a_1 x_1 + \cdots + a_n x_n\right] = \sum_{j=1}^{n} \sum_{k=1}^{n} a_j a_k \gamma(j-k)$$

  Also, the value at $h = 0$, namely

$$\gamma(0) = \text{E}\left[(x_t - \mu)^2\right]$$

  is the variance of the time series and the Cauchy-Schwarz inequality implies $|\gamma(h)| \leq \gamma(0)$
- The autocovariance function of a stationary series is symmetric around the origin

$$\gamma(h) = \gamma(-h)$$

  for all h because

$$\gamma((t+h)-t) = \text{Cov}\left[x_{t+h},\ x_t\right] = \text{Cov}\left[x_t,\ x_{t+h}\right] = \gamma(t-(t+h))$$

- Two time series, x_t and y_t, are said to be jointly stationary if they are each stationary and the cross-covariance function

$$\gamma_{xy}(h) = \text{Cov}\left[x_{t+h},\ y_t\right] = \text{E}\left[(x_{t+h} - \mu_x)(y_t - \mu_y)\right]$$

  is a function only of lag h
- The cross-correlation function (CCF) of jointly stationary time series x_t and y_t is defined as

$$\rho_{xy}(h) = \frac{\gamma_{xy}(h)}{\sqrt{\gamma_x(0)\gamma_y(0)}}$$

- The cross-correlation function is not generally symmetric about zero because $\text{Cov}\left[x_2,\ y_1\right]$ and $\text{Cov}\left[x_1,\ y_2\right]$ need not be the same
- Let two series, x_t and y_t be formed from the sum and difference of two successive values of a white noise process

$$x_t = w_t + w_{t-1} \text{ and } y_t = w_t - w_{t-1}$$

  where w_t are independent random variables with zero means and variance σ_w^2, then $\gamma_x(0) = \gamma_y(0) = 2\sigma_w^2$ and $\gamma_x(1) = \gamma_x(-1) = \sigma_w^2$, $\gamma_y(1) = \gamma_y(-1) = -\sigma_w^2$; also

$$\gamma_{xy}(1) = \text{Cov}\left[x_{t+1},\ y_t\right] = \text{Cov}\left[w_{t+1} + w_t,\ w_t - w_{t-1}\right] = \sigma_w^2$$

because only one term is nonzero; similarly, $\gamma_{xy}(0) = 0$ and $\gamma_{xy}(-1) = -\sigma_w^2$ and so

$$\rho_{xy}(h) = \begin{cases} 0 & \text{if } h = 0 \\ \frac{1}{2} & \text{if } h = 1 \\ -\frac{1}{2} & \text{if } h = -1 \\ 0 & \text{if } |h| \geq 2 \end{cases}$$

The autocovariance and cross-covariance functions only depend on the lag separation h and so the series are jointly stationary

- A linear process x_t is defined to be a linear combination of white noise variates w_t and is given by

$$x_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j w_{t-j}$$

where $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$

- For the linear process, the autocovariance function is given by

$$\gamma_x(h) = \sigma_w^2 \sum_{-\infty}^{\infty} \psi_{j+h}\psi_j$$

for $h \geq 0$

- The linear process is dependent on the future ($j < 0$), the present ($j = 0$) and the past ($j > 0$)

- A process $\{x_t\}$ is said to be a Gaussian process if the n-dimensional vectors $x = (x_{t_1}, x_{t_2}, \ldots, x_{t_n})'$, for every collection of distinct time points $t_1, t_2, \ldots, t_n$ and every positive integer n, have a multivariate normal distribution

- Defining the $n \times 1$ mean vector $\mathrm{E}[x] \equiv \mu \equiv (\mu_{t_1}, \mu_{t_2}, \ldots, \mu_{t_n})'$ and the $n \times n$ covariance matrix as $\mathrm{Var}[x] \equiv \Gamma = \{\gamma(t_i, t_j); i, j = 1, \ldots, n\}$, which is assumed to be positive definite, the multivariate normal density function can be written as

$$f(x) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\Gamma}} \exp\left\{-\frac{1}{2}(x-\mu)'\Gamma^{-1}(x-\mu)\right\}$$

for $x \in \mathbb{R}^n$

- If a Gaussian time series $\{x_t\}$ is weakly stationary, then μ_t is constant and $\gamma(t_i, t_j) = \gamma(|t_i - t_j|)$, so that the vector μ and the matrix Γ are independent of time; this implies that all the finite distributions of the series $\{x_t\}$ depend only on time lag and not on the actual times and hence the series must be strictly stationary

- A result called the Wold Decomposition states that a stationary non-deterministic time series is a causal linear process (but with $\sum \psi_j^2 < \infty$); a linear process need not be Gaussian but if a time series is Gaussian, then it is a causal linear process with $w_t \sim N(0, \sigma_w^2)$

- It is not enough for the marginal distributions to be Gaussian for the process to be Gaussian; it is easy to construct a situation where X and Y are normal but (X, Y) is not bivariate normal; e.g., let X and Z be independent normals and let $Y = Z$ is $XZ > 0$ and $Y = -Z$ if $XZ \leq 0$

1.5 Estimation of Correlation

- If a time series is stationary, the mean function $\mu_t = \mu$ is constant and can be estimated by the sample mean

$$\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$$

The standard error of the estimate is the square root of $\text{Var}[\bar{x}]$ which is given by

$$\begin{aligned}\text{Var}[\bar{x}] = \text{Var}\left[\frac{1}{n}\sum_{t=1}^{n} x_t\right] &= \frac{1}{n^2}\text{Cov}\left[\sum_{t=1}^{n} x_t,\ \sum_{s=1}^{n} x_s\right] \\ &= \frac{1}{n^2}(n\gamma_x(0) + (n-1)\gamma_x(1) + (n-2)\gamma_x(2) + \cdots + \gamma_x(n-1) \\ &+ (n-1)\gamma_x(-1) + (n-2)\gamma_x(-2) + \cdots + \gamma_x(1-n)) \\ &= \frac{1}{n}\sum_{h=-n}^{n}\left(1 - \frac{|h|}{n}\right)\gamma_x(h)\end{aligned}$$

If the process is white noise, this reduces to the familiar γ_x^2/n recalling that $\gamma_x(0) = \sigma_x^2$

- The sample autocovariance function is defined as

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_{t+h} - \bar{x})(x_t - \bar{x})$$

with $\hat{\gamma}(-h) = \hat{\gamma}(h)$ for $h = 0, 1, \ldots, n-1$

- The sum above is restricted over a range because x_{t+h} is not available for $t + h > n$

- The sample autocorrelation function is defined as

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$$

The sample autocorrelation function has a sampling distribution that allows assessment of whether the data comes from a completely random or white series or whether correlations are statistically significant at some lags

- Under general conditions (x_t is iid with finite fourth moment and so x_t is white Gaussian noise), if x_t is white noise, then for n large, the sample ACF, $\hat{\rho}_x(h)$, for $h = 1, 2, \ldots, H$, where H is fixed but arbitrary, is approximately normally distributed with zero mean and standard deviation given by

$$\sigma_{\hat{\rho}_x(h)} = \frac{1}{\sqrt{n}}$$

- A rough method of assessing whether peaks in $\hat{\rho}(h)$ are significant is by determining whether the observed peak is outside the interval $\pm 2/\sqrt{n}$ (or plus/minus two standard errors); for a white noise sequence, approximately 95% of the sample ACFs should be within these limits

- The estimators for the cross-covariance function $\gamma_{xy}(h)$ and the cross-correlation $\rho_{xy}(h)$ is given by the sample cross-covariance function

$$\hat{\gamma}_{xy}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_{t+h} - \bar{x})(y_t - \bar{y})$$

  where $\hat{\gamma}_{xy}(-h) = \hat{\gamma}_{yx}(h)$ determines the function for negative lags and the sample cross-correlation function

$$\hat{\rho}_{xy}(h) = \frac{\hat{\gamma}_{xy}(h)}{\sqrt{\hat{\gamma}_x(0)\hat{\gamma}_y(0)}}$$

- The sample cross-correlation function can be examined graphically as a function of lag h to search for leading or lagging relations in the data

- The large sample distribution of $\hat{\rho}_{xy}(h)$ is normal with mean zero and

$$\sigma_{\hat{\rho}_{xy}} = \frac{1}{\sqrt{n}}$$

  if at least one of the processes is independent white noise

- The autocorrelation and cross-correlation functions are useful for analyzing the joint behavior of two stationary series whose behavior may be related in some unspecified way

1.6 Vector-Valued and Multidimensional Series

- A vector time series, $x_t = (x_{t1}, x_{t2}, \ldots, x_{tp})'$, contains as its components p univariate time series; the $p \times 1$ column vector of the observed series is denoted as x_t and the row vector x_t' is its transpose

- For the stationary case, the $p \times 1$ mean vector $\mu = \mathrm{E}[x_t]$ is of the form $\mu = (\mu_{t1}, \mu_{t2}, \ldots, \mu_{tp})'$ and the $p \times p$ autocovariance matrix is

$$\Gamma(h) = \mathrm{E}[(x_{t+h} - \mu)(x_t - \mu)']$$

  where the elements of the matrix $\Gamma(h)$ are the cross-covariance functions

$$\gamma_{ij}(h) = \mathrm{E}[(x_{t+h,i} - \mu_i)(x_{tj} - \mu_j)]$$

  for $i, j = 1, \ldots, p$; because $\gamma_{ij}(h) = \gamma_{ji}(-h)$, $\Gamma(-h) = \Gamma'(h)$

- The sample autocovariance matrix of the vector series x_t is the $p \times p$ matrix of sample cross-covariances defined as

$$\hat{\Gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(x_t - \bar{x})'$$

  where $\bar{x} = \frac{1}{n} \sum_{t=1}^{n} x_t$ denotes the $p \times 1$ sample mean vector

- The symmetry property of the theoretical autocovariance extends to the sample autocovariance, which is defined for negative values by taking $\hat{\Gamma}(-h) = \hat{\Gamma}(h)'$

- Cases where an observed series may be indexed by more than time alone are defined as multidimensional process x_s where $s = (s_1, s_2, \ldots, s_r)'$ denotes the coordinate of the ith index

- The autocovariance function of a stationary multidimensional process, x_s, can be defined as a function of the multidimensional lag vector, $h = (h_1, h_2, \ldots, h_r)'$, as

$$\gamma(h) = \mathrm{E}\left[(x_{s+h} - \mu)(x_s - \mu)\right]$$

  where $\mu = \mathrm{E}\left[x_s\right]$ does not depend on the spatial coordinate s

- The multidimensional sample autocovariance function is defined as

$$\hat{\gamma}(h) = \frac{1}{S_1 S_2 \ldots S_r} \sum_{s_1} \sum_{s_2} \cdots \sum_{s_r} (x_{s+h} - \bar{x})(x_s - \bar{x})$$

  where $s = (s_1, s_2, \ldots, s_r)'$ and the range of summation for each argument is $1 \leq s_i \leq S_i - h_i$ for $i = 1, \ldots, r$; the mean is computed over the r-dimensional array

$$\bar{x} = \frac{1}{S_1 S_2 \ldots S_r} \sum_{s_1} \sum_{s_2} \cdots \sum_{s_r} x_{s_1, s_2, \ldots, s_r}$$

  where the arguments s_i are summed over $1 \leq s_i \leq S_i$

- The multidimensional sample autocorrelation function follows by taking the scaled ratio

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$$

2 Time Series Regression and Explanatory Data Analysis

2.1 Classical Regression in the Time Series Context

- Assume some output or dependent time series, x_t for $t = 1, \ldots, n$ is being influenced by a collection of possible inputs or independent series $z_{t1}, z_{t2}, \ldots, z_{tq}$ where the inputs are fixed and known; this relation is defined through the linear regression model

$$x_t = \beta_0 + \beta_1 z_{t1} + \beta_2 z_{t2} + \cdots + \beta_q z_{tq} + w_t$$

  where $\beta_0, \beta_1, \ldots, \beta_q$ are unknown fixed regression coefficients and $\{w_q\}$ is a random error or noise process consisting of independent and identically distributed (iid) normal variables with mean 0 and variance σ_w^2

- In ordinary least squares (OLS), the error sum of squares is minimized

$$Q = \sum_{t=1}^{n} w_t^2 = \sum_{t=1}^{n} (x_t - [\beta_0 + \beta_1 z_t])^2$$

  with respect to β_i for $i = 0, 1$

- The OLS estimates of the coefficients are obtained by evaluating $\partial Q / \partial \beta_i = 0$ for $i = 0, 1$ and solving for the βs

$$\hat{\beta}_1 = \frac{\sum_{t=1}^{n} (x_t - \bar{x})(z_t - \bar{z})}{\sum_{t=1}^{n} (z_t - \bar{z})^2} \text{ and } \hat{\beta}_0 = \bar{x} - \hat{\beta}_1 \bar{z}$$

  where $\bar{x} = \sum_t x_t / n$ and $\bar{z} = \sum_t z_t / n$ are the respective sample means

- The multiple linear regression model can be conveniently written in a more general notation by defining the column vectors $z_t = (1, z_{t1}, z_{t2}, \ldots, z_{tq})'$ and $\beta = (\beta_0, \beta_1, \ldots, \beta_q)'$ where $'$ denotes transpose; then the linear regression model equation can be rewritten as

$$x_t = \beta_0 + \beta_1 z_{t1} + \cdots + \beta_q z_{tq} + w_t = \beta' z_t + w_t$$

  where $w_t \sim N(0, \sigma_w^2)$

- OLS estimation finds the coefficient vector β that minimizes the error sum of squares

$$Q = \sum_{t=1}^{n} w_t^2 = \sum_{t=1}^{n} (x_t - \beta' z_t)^2$$

  with respect to $\beta_0, \beta_1, \ldots, \beta_q$; by setting $\sum_{t=1}^{n} (x_t - \hat{\beta}' z_t) z_t' = 0$, the normal equations are obtained

$$\left(\sum_{t=1}^{n} z_t z_t' \right) \hat{\beta} = \sum_{t=1}^{n} z_t x_t$$

If the term inside the parenthesis is non-singular, then the least squares estimate of β is

$$\hat{\beta} = \left(\sum_{t=1}^{n} z_t z_t' \right)^{-1} \sum_{t=1}^{n} z_t x_t$$

- The minimized error sum of squares, SSE, can be written as

$$\text{SSE} = \sum_{t=1}^{n} (x_t - \hat{\beta}' z_t)^2$$

- The OLS estimators are unbiased, meaning $\text{E}\left[\hat{\beta}\right] = \beta$ and have the smallest variance within the class of linear unbiased estimators
- If the errors w_t are normally distributed, $\hat{\beta}$ is the maximum likelihood estimator for β and is normally distributed with

$$\text{Cov}\left[\hat{\beta}\right] = \frac{\sigma_w^2}{\sum_{t=1}^{n} z_t z_t'}$$

- An unbiased estimator for the variance σ_w^2 is

$$s_w^2 = \text{MSE} = \frac{\text{SSE}}{n - (q+1)}$$

where MSE is the mean squared error

- Under the normal assumption,

$$t = \frac{\hat{\beta}_i - \beta_i}{s_w \sqrt{c_i i}}$$

has the t-distribution with $n-(q+1)$ degrees of freedom and c_{ii} denotes the ith diagonal element of $C = \left(\sum_{t=1}^{n} z_t z_t'\right)^{-1}$; this result is often used for individual tests of the null hypothesis $H_0 : \beta_i = 0$ for $i = 1, \ldots, q$

- Suppose a proposed model specifies that only a subset $r < q$ independent variables, $z_{t,1:r} = \{z_{t1}, z_{t2}, \ldots, z_{tr}\}$ is influencing the dependent variable x_t, then the reduced model is

$$x_t = \beta_0 + \beta_1 z_{t1} + \cdots + \beta_r z_{tr} + w_t$$

where $\beta_1, \beta_2, \ldots, \beta_r$ are a subset of coefficients of the original q variables; the null hypothesis in this case is $H_0 : \beta_{r+1} = \cdots = \beta_q = 0$; this reduced model can be tested against the full model by computing the error sums of squares under the two models using the F-statistic

$$F = \frac{(\text{SSE}_r - \text{SSE})/(q-r)}{\text{SSE}/(n-q-1)} = \frac{\text{MSR}}{\text{MSE}}$$

where SSE_r is the error sum of squares under the reduced model; under the null hypothesis, this has a central F-distribution with $q-r$ and $n-q-1$ degrees of freedom when the reduced model is the correct model

- Analysis of Variance for Regression

Source	df	Sum of Squares	Mean Square	F
$z_{t,r+1:q}$	$q-r$	$\text{SSR} = \text{SSE}_r - \text{SSE}$	$\text{MSR} = \text{SSR}/(q-r)$	$F = \frac{\text{MSR}}{\text{MSE}}$
Error	$n-(q+1)$	SSE	$\text{MSE} = \text{SSE}/(n-q-1)$	

- The difference in the numerator is often called the regression sum of squares (SSR); the null hypothesis is rejected at level α if $F > F^{q-r}_{n-q-1}(\alpha)$, the $1-\alpha$ percentile of the F distribution with $q-r$ numerator and $n-q-1$ denominator degrees of freedom

- A special case is the null hypothesis $H_0 : \beta_1 = \cdots = \beta_q = 0$; here the model is $x_t = \beta_0 + w_t$; the proportion of variation accounted for by all the variables is measured using

$$R^2 = \frac{\text{SSE}_0 - \text{SSE}}{\text{SSE}_0}$$

where the residual sum of squares under the reduced model is $\text{SSE}_0 = \sum_{t=1}^{n}(x_t - \bar{x})^2$

- SSE_0 is the sum of squared deviations from the mean $\bar{x}$ and is otherwise known as the adjusted total sum of squares; the measure R^2 is called the coefficient of determination

- In a process called stepwise multiple regression, variables are added or deleted from the model by testing various models against one another using the F test until a set of useful variables is found

- Suppose a normal regression model has k coefficients and the maximum likelihood estimator for the variance is denoted as

$$\hat{\sigma}^2_k = \frac{\text{SSE}(k)}{n}$$

where $\text{SSE}(k)$ denotes the residual sum of squares under the model with k regression coefficients; then the goodness of fit for this model can be measured by balancing the error of the fit against the number of parameters in the model using the AIC

- Akaike's Information Criterion (AIC)

$$\text{AIC} = \log \hat{\sigma}^2_k + \frac{n+2k}{n}$$

where $\hat{\sigma}^2_k$ is given as above and k is the number of parameters in the model

- The value of k yielding the minimum AIC specifies the best model; the idea is roughly that minimizing $\hat{\sigma}^2_k$ would be a reasonable objective, except that it decreases monotonically as k increases and so, the error variance is penalized by a term proportional to the number of parameters

- To account for bias based on small-sample distributional results or the linear regression model, use AIC_c

$$\text{AIC}_c = \log \hat{\sigma}^2_k + \frac{n+k}{n-k-2}$$

where n is the sample size

- A correction term can only be derived using Bayesian ideas
- Bayesian Information Criterion (BIC)

$$\text{BIC} = \log \hat{\sigma}_k^2 + \frac{k \log n}{n}$$

- The penalty term in BIC much larger than in AIC and so BIC tends to choose smaller models
- BIC does well at getting the correct order in large samples whereas AIC_c tends to be superior in smaller samples where the relative number of parameters is large
- Two additional measures for fitting regressional models are adjusted R-squared (which is essentially s_w^2) and Mallows C_p

2.2 Explanatory Data Analysis

- In general, it is necessary for time series data to be stationary so that averaging lagged products over time will be a sensible thing to do; with time series data, it is the dependence between the values of the series that is important to measure; autocorrelations must be able to be estimated with precision
- To achieve any meaningful statistical analysis of time series data, it is important that the mean and the autocovariance functions satisfy the conditions of stationarity (for at least some reasonable stretch of time)
- The easiest form of nonstationarity to work with is the trend stationary model where the process has stationary behavior around a trend

$$x_t = \mu_t + y_t$$

  where x_t are the observations, μ_t denotes the trend and y_t is a stationary process; in these models, strong trend will obscure the behavior of the stationary process y_t; hence it is advantageous to remove the trend as a first step in an explanatory analysis of such time series; the steps involved are to obtain a reasonable estimate of the trend component, $\hat{\mu}_t$, and then work with the residuals

$$\hat{y}_t = x_t - \hat{\mu}_t$$

- A random walk with drift model is a good model for trend; that is, rather than modeling trend as fixed, model trend as a stochastic component using the random walk with drift model

$$\mu_t = \delta + \mu_{t-1} + w_t$$

  where w_t is white noise and is independent of y_t; if the appropriate model is the trend stationary model, then differencing the data, x_t, yields a stationary process

$$\begin{aligned} x_t - x_{t-1} &= (\mu_t + y_t) - (\mu_{t-1} + y_{t-1}) \\ &= \delta + w_t + y_t - y_{t-1} \end{aligned}$$

and $z_t = y_t - y_{t-1}$ is stationary because y_t is stationary

$$\begin{aligned}\gamma_z(h) &= \text{Cov}\left[z_{t+h},\ z_t\right] = \text{Cov}\left[y_{t+h} - y_{t+h-1},\ y_t - y_{t-1}\right] \\ &= 2\gamma_y(h) - \gamma_y(h+1) - \gamma_y(h-1)\end{aligned}$$

which is independent of time

- One advantage of differencing over detrending to remove trend is that no parameters are estimated in the differencing operation; one disadvantage is that differencing does not yield an estimate of the stationary process y_t; if an estimate of y_t is important, then detrending may be more appropriate; if the goal is to coerce the data to stationarity, then differencing may be more appropriate

- Differencing is also a viable tool if the trend is fixed; if $\mu_t = \beta_0 + \beta_1 t$ in the trend stationary model, differencing the data produces stationarity

$$x_t - x_{t-1} = (\mu_t + y_t) - (\mu_{t-1} + y_{t-1}) = \beta_1 + y_t - y_{t-1}$$

- Because differencing plays a big role in time series analysis, it has its own notation; the first difference is denoted as

$$\nabla x_t = x_t - x_{t-1}$$

The first difference eliminates a linear trend; a second difference (the difference of ∇x_t) can eliminate a quadratic trend, and so on

- The backshift operator is defined as

$$Bx_t = x_{t-1}$$

It can be extended to powers $B^2 x_t = B(Bx_t) = Bx_{t-1} = x_{t-2}$ and so on; thus

$$B^k x_t = x_{t-k}$$

- The idea of an inverse operator can also be given if $B^{-1}B = 1$, so that

$$x_t = B^{-1}Bx_t = B^{-1}x_{t-1}$$

This means B^{-1} is the forward shift operator

- The first difference can then be rewritten as

$$\nabla x_t = (1 - B)x_t$$

whereas the second difference is

$$\nabla^2 x_t = (1 - B)^2 x_t = (1 - 2B + B^2)x_t = x_t - 2x_{t-1} + x_{t-2}$$

by the linearity of the operator; this notion can be extended as such further on

- Differences of order d are defined as

$$\nabla^d = (1 - B)^d$$

where the RHS can be expanded algebraically to evaluate for higher integer values of d

- The first difference is an example of a linear filter applied to eliminate a trend; other filters, formed by averaging values near x_t, can produce adjusted series that eliminate other kinds of unwanted fluctuations

- An alternative to differencing is fractional differencing, where the notion of the difference operator is extended to fractional powers $-0.5 < d < 0.5$ which still define stationary processes

- When there are obvious aberrations that contribute nonstationary and nonlinear behavior in observed time series, transformations can be useful to equalize the variability over the length of a single series

- A useful transformation is

$$y_t = \log x_t$$

which tends to suppress larger fluctuations that occur over portions of the series where the underlying values are larger

- Other possibilities are power transformations in the Box-Cox family of the form

$$y_t = \begin{cases} (x_t^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log x_t & \text{if } \lambda = 0 \end{cases}$$

There are methods for choosing the power λ

- Often transformations are also used to improve the approximation to normality or to improve linearity in predicting the value of one series from another

- Another preliminary data processing technique for visualizing the relations between series at different lags is scatterplot matrices which allows examining scatterplots of y_t versus x_{t-h} to find linear correlations and predictability

- To check for nonlinear relations, it is convenient to display a lagged scatterplot matrix that displays values of the series, x_t, on the vertical axis plotted against x_{t-h} on the horizontal axis; the sample autocorrelations are locally weighed scatterplot smoothing (lowess) lines that can be used to help discover any nonlinearities

- It can also be useful to look at values of one series plotted against another series at various lags, x_{t-h}, to look for potential nonlinear relations between the two series

2.3 Smoothing in the Time Series Analysis

- Using a moving average to smooth white noise is useful in discovering certain traits in a time series, such as long term trend and seasonal components
- If x_t represents the observations, then

$$m_t = \sum_{j=-k}^{k} a_j x_{t-j}$$

where $a_j = a_{-j} \geq 0$ and $\sum_{j=-k}^{k} a_j = 1$ is a symmetric moving average of the data

- Kernel smoothing is a moving average smoother that uses a weight function, or kernel, to average the observations; m_t is now

$$m_t = \sum_{i=1}^{n} w_i(t) x_i$$

where

$$w_i(t) = K\left(\frac{t-i}{b}\right) / \sum_{j=1}^{n} K\left(\frac{t-j}{b}\right)$$

are the weights and $K(\cdot)$ is a kernel function; this estimator is often called the Nadaraya-Watson estimator

- Typically, the normal function is used for the kernel function

$$K(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

- Another approach to smoothing a time plot is nearest neighbor regression which is based on k-nearest neighbors regression where one uses only the data

$$\{x_{t-k/2}, \ldots, x_t, \ldots, x_{t+k/2}\}$$

to predict x_t via regression and then sets $m_t = \hat{x}_t$

- Lowess is a method of smoothing that is complex but close to nearest neighbor regression; first, a certain proportion of nearest neighbors to x_t are included in a weighting scheme; values closer to x_t in time get more weight; then, a robust weighted regression is used to predict x_t and obtain the smoothed values m_t; the larger the fraction of nearest neighbors included, the smoother the fit will be
- Another way to smooth data is to fit a polynomial regression in terms of time

$$m_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \ldots$$

and then fitting m_t via OLS

- An extension of polynomial regression is to first divide time $t = 1, \ldots, n$ into k intervals, $[t_0 = 1, t_1]$, $[t_1 + 1, t_2]$, $\ldots$, $[t_{k-1}, t_k = n]$; the values $t_0, t_1, \ldots, t_k$ are called knots; then in each interval, one fits a polynomial regression; typically, the order k is 3 and so is called cubic splines
- A related method is called smoothing splines which minimizes a compromise between the fit and the degree of smoothness given by

$$\sum_{t=1}^{n} [x_t - m_t]^2 + \lambda \int (m_t'')^2 \, dt$$

  where m_t is a cubic spline with a knot at each t and primes denote differentiation; the degree of smoothness is controlled by $\lambda > 0$ where the larger the value of λ is, the smoother the fit is
- If $\lambda = 0$, $m_t = x_t$ and so the fit is not smooth at all; if $\lambda = \infty$, the fit is completely smooth; thus λ is seen as a trade-off between linear regression (completely smooth) and the data itself (no smoothness)
- In addition to smoothing time plots, smoothing techniques can be applied to smoothing a time series as a function of another time series

3 ARIMA Models

3.1 Autoregressive Moving Average Models

3.2 Difference Equations

3.3 Autocorrelation and Partial Autocorrelation

3.4 Forecasting

3.5 Estimation

3.6 Integrated Models for Nonstationary Data

3.7 Building ARIMA Models

3.8 Regression with Autocorrelated Errors

3.9 Multiplicative Seasonal ARIMA Models

4 Spectral Analysis and Filtering

4.1 Cyclical Behavior and Periodicity

4.2 The Spectral Density

4.3 Periodogram and Discrete Fourier Transform

4.4 Nonparametric Spectral Estimation

4.5 Parametric Spectral Estimation

4.6 Multiple Series and Cross-Spectra

4.7 Linear Filters

4.8 Lagged Regression Models

4.9 Signal Extraction and Optimum Filtering

4.10 Spectral Analysis and Multidimensional Series

5 Additional Time Domain Topics

5.1 Long Memory ARMA and Fractional Differencing

5.2 Unit Root Testing

5.3 GARCH Models

5.4 Threshold Models

5.5 Lagged Regression and Transfer Function Modeling