STAT 443 - Forecasting

Cameron Roopnarine

Last updated: January 19, 2021

Contents

Contents **1**

1 Characteristics of Time Series **2**
1.1 What is a time series? . . . 2
1.2 Basic Principles of Forecasting . . . 4
1.3 Definitions of Stationary . . . 6
1.4 White Noise and Stationary Examples . . . 7
1.5 Weak versus Strong Stationary . . . 9
1.6 † Theoretical L2 Framework for Time Series . . . 11
1.6.1 Useful Tools for Time Series . . . 12
1.7 Signal and Noise Models . . . 12
1.8 Time Series Differencing . . . 14

2 Time Series Regression **18**
2.1 Autocorrelation and Empirical Autocorrelation . . . 18

Chapter 1

Characteristics of Time Series

1.1 What is a time series?

In classical statistics, we normally consider $X_1, \dots, X_n \in \mathbf{R}^p$, a **simple random sample**.

In particular,

(1) $X_1, \dots, X_n$ are i.i.d. (independent and identically distributed)

(2) $X_i \sim F_\theta$ which is a common distribution characterized by θ.

Examples:

1. $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, and we wish to estimate and perform inference on μ and σ^2.
2. $X_i = \begin{bmatrix} Y_i \\ Z_i \end{bmatrix}$ where Y_i is a dependent variable, and Z_i is an independent variable. Perhaps we happen to observe Y_i and Z_i in pairs, and we stop at a model:

$$Y_i = \beta^\top Z_i + \varepsilon_i \quad \text{where } \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$$

REMARK 1.1.1

The relationship between Y_i and Z_i doesn't depend on i, it only depends through the common parameter β, and it assumes that ε has fixed variance for each i.

3. In such settings, one is typically interested in:
   (a) Prediction: how based on the data, can we predict how these variables would behave in the future?
   (b) Inference: how do we use the data to try to estimate and understand better the underlying mechanism which generates the data? For example, a linear model or simple Gaussian model.

DEFINITION 1.1.2: Time Series

We say $X_1, \dots, X_T$ is an (observed) **time series** of length T if X_t denotes an observation obtained at time t.
In particular, the observations are ordered in time.

DEFINITION 1.1.3: Real-valued

If $X_t \in \mathbf{R}$, we say $X_1, \dots, X_T$ is a **real-valued** or **scalar** time series.

DEFINITION 1.1.4: Multivariate

If $X_t \in \mathbf{R}^p$, we say $X_1, \dots, X_T$ is a **multivariate** or **vector-valued** time series.

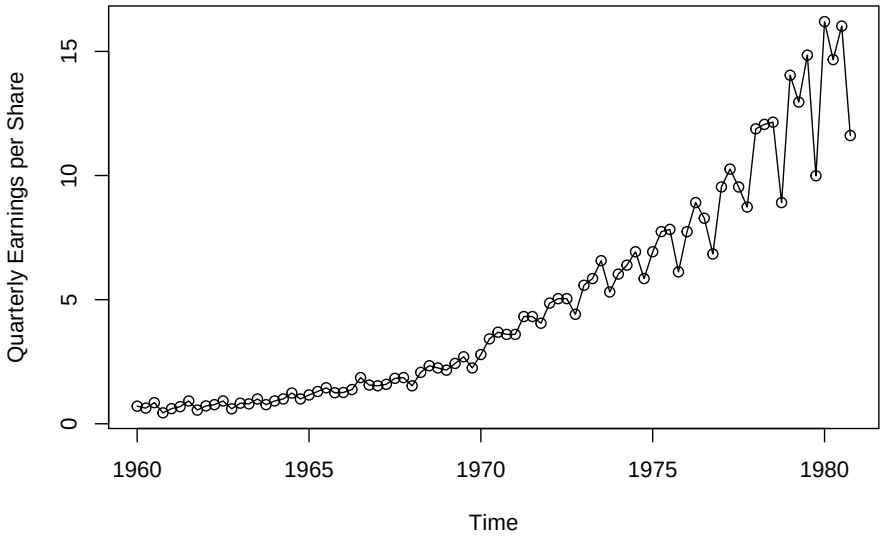

Figure 1.1: Quarterly Johnson and Johnson Earnings

```
# Figure 1.1
plot(jj, type = "o", ylab = "Quarterly Earnings per Share")
```

Observe that in Figure 1.1:

- The earnings are steadily increasing over time.
- There is heterogeneity in the variance over time.

With time series data, we are typically concerned with the same goals as in classical statistics (prediction and inference). However, in contrast, with time series, the data often exhibit:

(1) Heterogeneity

- Time trends $\rightarrow \mathbb{E}[X_t] \neq \mathbb{E}[X_{t+h}]$
- Heteroskedasticity $\rightarrow \mathbb{V}(X_t) \neq \mathbb{V}(X_{t+h})$

In classical statistics, it's assumed that all the observations have the same distribution which is clearly not the case in time series.

(2) Serial Dependence (Serial Correlation)

- Observations that are temporally close appear to depend on each other.

In classical statistics, each successive observation is assumed to be independent which is clearly not the case in time series.

```
# Figure 1.2
plot(gtemp, type = "o", ylab = "Global Temperature Deviations")
```

Observe that in Figure 1.2:

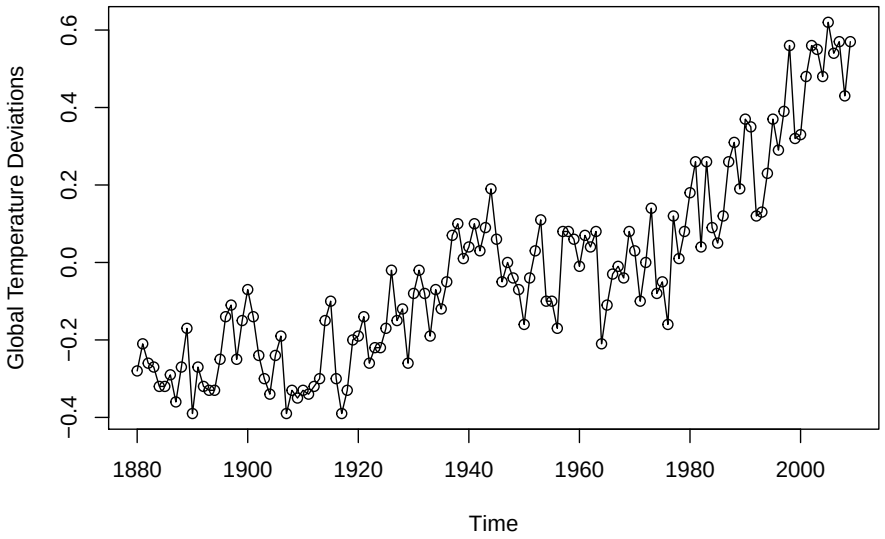

Figure 1.2: x_t is the deviation of global mean yearly temperature from the mean computed from 1951–1980

- The global temperature is steadily increasing over time.
- There is heterogeneity in the mean over time.
- Heterogeneity in the variance over time is not very apparent.
- Serial dependence occurs.

DEFINITION 1.1.5: Time series (Formal Definition)

We say $\{X_t\}_{t\in\mathbf{Z}}$ is a **time series** if $\{X_t : t \in \mathbf{Z}\}$ is a stochastic process indexed by **Z**.

In other words, there is a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ so that for all $X_t : \Omega \to \mathbf{R}$ is a random variable.

In relation to the original definition, we say $X_1, \dots, X_T$ is an **observed stretch** (**realization**, **simple path**) of length T from $\{X_t\}_{t\in\mathbf{Z}}$.

Formally speaking, we think of a time series as being a little snippet of one long sample path the stochastic process for which would characterize all the serial dependence, time trends, and heteroskedasticity, that exist within a time series.

1.2 Basic Principles of Forecasting

Consider a time series of length T, namely $X_1, \dots, X_T$. Based on $X_1, \dots, X_T$, we would like to produce a "best guess" for X_{T+h}:

$$\hat{X}_{T+h} = \hat{X}_{T+h|T} = f_h(X_T, \dots, X_1)$$

DEFINITION 1.2.1: Forecast, Horizon

For $h \geq 1$, our "best guess"

$$\hat{X}_{T+h} = f_h(X_T, \dots, X_1)$$

is called a **forecast** of X_{T+h} at **horizon** h.

There are two primary goals in forecasting:

- Goal 1: Choose f_n "optimally." Normally, we or the practitioner have some measure, say $L(\cdot, \cdot)$, in mind for determining how "close" $\hat{X}_{T+h}$ is to the true value, X_{T+h}. We then wish to choose f_h so that $L(X_{T+h}, f_h(X_T, \dots, X_1))$ is minimized.

  EXAMPLE 1.2.2

  Most common measure $L(\cdot, \cdot)$ is mean-squared error (MSE), where

  $$L(X, Y) = \mathbb{E}[(X - Y)^2]$$

- **Goal 2**: Quantify the uncertainty in the forecast. This entails providing some description of how close we expect $\hat{X}_{T+h}$ to be to X_{T+h}.

  EXAMPLE 1.2.3: Why is it important to quantify uncertainty?

  Suppose every minute, we flip a coin and denote
  - $H \to 1$
  - $T \to -1$
  - $X_t =$ outcome in minute t, where $t = 1, \dots, T$.

  This produces a time series of length T, which is a random sequence of (1)'s and (-1)'s. Note $\mathbb{E}[X_t] = 0$. So, if we wish to forecast for $h \geq 1$, consider $\hat{X}_{T+h} = f(X_T, \dots, X_1)$

  $$\begin{aligned} L(X_{T+h}, \hat{X}_{T+h}) &= \mathbb{E}[(X_{T+h} - \hat{X}_{T+h})^2] \\ &= \mathbb{E}[X_{T+h}^2] + \mathbb{E}[\hat{X}_{T+h}^2] - 2\mathbb{E}[X_{T+h}\hat{X}_{T+h}] \\ &= \mathbb{E}[X_{T+h}^2] + \mathbb{E}[\hat{X}_{T+h}^2] - 2\mathbb{E}[X_{T+h}]\mathbb{E}[\hat{X}_{T+h}] \\ &= \mathbb{E}[X_{T+h}^2] + \mathbb{E}[\hat{X}_{T+h}^2] \end{aligned}$$

  Furthermore, note that $\mathbb{E}[X_{T+h}^2] = \mathbb{V}(X_t)$ since $\mathbb{E}[X_{T+h}] = 0$.
  We can write $\mathbb{E}[X_{T+h}\hat{X}_{T+h}] = \mathbb{E}[X_{T+h}]\mathbb{E}[\hat{X}_{T+h}]$ since $\hat{X}_{T+h}$ is a function of the data $X_T, \dots, X_1$, and hence independent of X_{T+h}.
  We can minimize this by taking $\hat{X}_{T+h} = 0$. There's nothing "wrong" with this forecast. But ideally, we would also be able to say that the sequence appears to be random, and that we don't expect this forecast to be close to the actual value.
  Furthermore, for this basic reason, one can always argue that any forecast that's not accompanied with some type of quantification of how close we expect the forecast to be, is at very least hard to interpret; at worst, meaningless because it doesn't describe the accuracy for which we expect the forecast to perform.

How can we quantify the uncertainty in forecasting?

Ideal: The predictive distribution:

$$X_{T+h} \mid X_T, \dots, X_1$$

Excellent: Predictive intervals/sets. For some $\alpha \in (0, 1)$ find I_α so that

$$\mathbb{P}(X_{T+h} \in I_\alpha \mid X_T, \dots, X_1) = \alpha$$

($\alpha = 0.95$, for example). Often such intervals take the form

$$I_\alpha = (\hat{X}_{T+h} - \hat{\sigma}_h, \hat{X}_{T+h} + \hat{\sigma}_h)$$

Concluding remarks:

1. Estimating predictive distribution leads one towards estimating the joint distribution of

$$X_{T+h}, X_T, \dots, X_1$$

   For example, ARMA and ARIMA models.

2. It is important that we acknowledge that some things cannot be predicted!

"It is tough to make predictions, especially about the future."—Yogi Berra

1.3 Definitions of Stationary

Given a time series $X_1, \dots, X_T$, we are frequently interested in estimating the joint distribution of

$$X_{T+h}, X_T, \dots, X_1$$

which is useful for forecasting and inference.

The joint distribution is a feature of the process $\{X_t\}_{t \in \mathbf{Z}}$

$$X_1, \dots, X_T \xrightarrow[\text{infer}]{} \{X_t\}_{t \in \mathbf{Z}}$$

- $X_1, \dots, X_T$: Observed data.
- $\{X_t\}_{t \in \mathbf{Z}}$: Stochastic process.

Worst case: $X_t \sim F_t$, where F_t is a *changing* function of t. If so, it is hard to pool the data $X_1, \dots, X_T$, to estimate F_t. If **serial dependence** occurs; that is, if the distribution of (X_t, X_{t+h}) depends strongly on t, then we have a similar problem in estimating e.g. $\mathsf{Cov}(X_t, X_{t+h})$.

DEFINITION 1.3.1: Strictly stationary

We say that a time series $\{X_t\}_{t \in \mathbf{Z}}$ is **strictly stationary** (**strongly stationary**) if for each $k \geq 1$, $i_1, \dots, i_k, h \in \mathbf{Z}$,

$$(X_{i_1}, \dots, X_{i_k}) \overset{\mathrm{d}}{=} (X_{i_1+h}, \dots, X_{i_k+h})$$

If we look at the k-dimensional joint distribution $(X_{i_1}, \dots, X_{i_k})$ of the series at points $i_1, \dots, i_k$, then strict stationary means this is shift-invariant. That is, shifting the window on which you view the data, does not change its distribution. This implies that if $F_t =$ CDF of X_t, then $F_t = F_{t+h} = F$ that is, all variables have a common distribution function.

DEFINITION 1.3.2: Mean function

For a time series $\{X_t\}_{t \in \mathbf{Z}}$, with $\mathbb{E}[X_t^2] < \infty$ for all $t \in \mathbf{Z}$, we denote the **mean function** of the time series as

$$\mu_t = \mathbb{E}[X_t]$$

DEFINITION 1.3.3: Autocovariance function, Lag

The **autocovariance** function of the time series $\{X_t\}_{t \in \mathbf{Z}}$ is defined as

$$\gamma(t, s) = \mathbb{E}[(X_t - \mu_t)(X_s - \mu_s)] = \mathsf{Cov}(X_t, X_s)$$

DEFINITION 1.3.4: Weakly stationary, Lag

We say that a time series $\{X_t\}_{t\in\mathbf{Z}}$ is **weakly stationary** if $\mathbb{E}[X_t] = \mu$ (does not depend on t), and if

$$\gamma(t,s) = f(|t-s|)$$

that is, $\gamma(t,s)$ is a function of $|t-s|$. In this case, we usually write

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

and we call the input h the **lag** parameter.

Additional terminology:

- The property when $\mathbb{E}[X_t] = \mu$ does not depend on t is often called the **first order stationary**.
- The property when $\gamma(t,s) = f(|t-s|)$ only depends on the lag $|t-s|$ is called the **second order stationary**.
- For a second order stationary process,

$$\begin{aligned}\gamma(h) &= \mathsf{Cov}(X_t, X_{t+h}) \\ &= \mathsf{Cov}(X_{t-h}, X_{t-h+h}) \qquad\qquad t \to (t-h) \\ &= \mathsf{Cov}(X_t, X_{t-h}) \\ &= \gamma(-h)\end{aligned}$$

  Since $\gamma(h) = \gamma(-h)$, we normally, we only record $\gamma(h)$ for $h \geq 0$.

1.4 White Noise and Stationary Examples

DEFINITION 1.4.1: Strong white noise

We say $\{X_t\}_{t\in\mathbf{Z}}$ is a **strong white noise** if $\mathbb{E}[X_t] = 0$ and the $\{X_t\}_{t\in\mathbf{Z}}$ are i.i.d.

DEFINITION 1.4.2: Weak white noise

We say $\{X_t\}_{t\in\mathbf{Z}}$ is a **weak white noise** if $\mathbb{E}[X_t] = 0$ and

$$\gamma(t,s) = \mathsf{Cov}(X_t, X_s) = \begin{cases}\sigma^2 & |t-s| = 0 \\ 0 & |t-s| > 0\end{cases}$$

DEFINITION 1.4.3: Gaussian white noise

We say $\{X_t\}_{t\in\mathbf{Z}}$ is a **Gaussian white noise** if $X_t \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2)$.

```
# Figure 1.3
plot.ts(rnorm(500), main = "Gaussian White Noise", ylab = "w")
```

Figure 1.3 is a Gaussian *white* noise series. **White** comes from spectral analysis, in which a white noise series shares the same spectral properties as white light: all periodicities occur with equal strength.

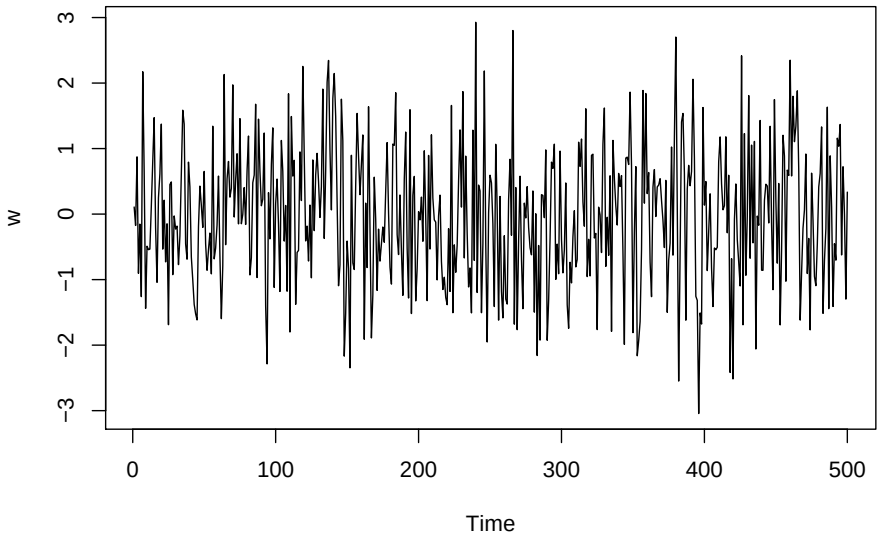

Figure 1.3: Gaussian White Noise of Length 500

EXAMPLE 1.4.4

Suppose $\{W_t\}_{t\in\mathbf{Z}}$ is a strong white noise, then $\mathbb{E}[W_t] = 0$; that is, it doesn't depend on t.

$$\gamma(t,s) = \mathsf{Cov}(W_t, W_s) = \mathbb{E}[W_t W_s] = \begin{cases} \sigma_W^2 & |t-s| = 0 \\ 0 & |t-s| > 0 \end{cases}$$

only depends on $|t-s|$.
$\{W_t\}_{t\in\mathbf{Z}}$ is **weakly stationary**. Furthermore, $\{W_t\}_{t\in\mathbf{Z}}$ is **strictly stationary**. Let $k \geq 1$ with $i_1 < \cdots i_k$ and $h \in \mathbf{Z}$, then

$$\begin{aligned} \mathbb{P}(W_{i_1} \leq t_1, \ldots, W_{i_k} \leq t_k) &= \prod_{j=1}^{k} \mathbb{P}(W_{i_j} \leq t_j) && \text{independence} \\ &= \prod_{j=1}^{k} \mathbb{P}(W_{i_j+h} \leq t_j) \\ &= \mathbb{P}(W_{i_1+h} \leq t_1, \ldots, W_{i_k+h} \leq t_k) \end{aligned}$$

EXAMPLE 1.4.5

Suppose $\{W_t\}_{t\in\mathbf{Z}}$ is a strong white noise. Define $X_t = W_t + \theta W_{t-1}$ for $\theta \in \mathbf{R}$. Since $\{W_t\}_{t\in\mathbf{Z}}$ is a strong white noise, we have $\mathbb{E}[W_t] = 0$ for all t, and so we have $\mathbb{E}[X_t] = \mathbb{E}[W_t + \theta W_{t-1}] = \mathbb{E}[W_t] + \theta\mathbb{E}[W_{t-1}] = 0$ which is first order stationary.

$$\gamma(t,s) = \mathsf{Cov}(X_t, X_s) = \begin{cases} (1+\theta^2)\sigma_W^2 & |t-s| = 0 \\ \theta\sigma_W^2 & |t-s| = 1 \\ 0 & |t-s| > 0 \end{cases}$$

We obtain these calculations as follows:

- $|t-s|=0$.

$$\mathbb{E}[(W_t+\theta W_{t-1})^2]=\mathbb{E}[W_t^2]+\theta^2\mathbb{E}[W_{t-1}^2]+2\mathbb{E}[\theta W_t W_{t-1}]=(1+\theta^2)\sigma_W^2$$

- $t=s+1$ (or $s=t+1$).

$$\mathbb{E}[(W_{s+1}+\theta W_s)(W_s+\theta W_{s-1})]=\theta\mathbb{E}[W_s^2]=\theta\sigma_W^2$$

since W_{s+1} is independent of W_s and W_{s-1}. The calculation is easy to verify.

- $|t-s|>1$. $W_t+\theta W_{t-1}$ is independent of $W_s+\theta W_{s-1}$.

$\{X_t\}_{t\in\mathbf{Z}}$ is also strictly stationary. Suppose $k\geq 1$, $i_1,\dots,i_k,h\in\mathbf{Z}$ with $i_1<\cdots<i_k$, then

$$\begin{aligned}\mathbb{P}(X_{i_1}\leq t_1,\dots,X_{i_k}\leq t_k)&=\mathbb{P}(W_{i_1}+\theta W_{i_1-1}\leq t_1,\dots,W_{i_k}+\theta W_{i_k-1}\leq t_k)\\&=\mathbb{P}\left(\begin{bmatrix}W_{i_1-1}\\W_{i_1}\\\vdots\\W_{i_k}\end{bmatrix}\in B\right)\\&=\mathbb{P}\left(\begin{bmatrix}W_{i_1-1+h}\\\vdots\\W_{i_k+h}\end{bmatrix}\in B\right)\\&=\mathbb{P}(X_{i_1+h}\leq t_1,\dots,X_{i_k+h}\leq t_k)\end{aligned}$$

where B is some subset of $\mathbf{R}^{i_k-i_1+1}$, and hence is shift-invariant.

DEFINITION 1.4.6: Bernoulli shift

Suppose $\{\varepsilon_t\}_{t\in\mathbf{Z}}$ is a strong white noise. If $X_t=g(\varepsilon_t,\varepsilon_{t-1},\dots)$ for some function $g:\mathbf{R}^\infty\to\mathbf{R}$, we say that $\{X_t\}_{t\in\mathbf{Z}}$ is a **Bernoulli shift**.

THEOREM 1.4.7

If $\{X_t\}_{t\in\mathbf{Z}}$ is a Bernoulli shift, then $\{X_t\}_{t\in\mathbf{Z}}$ is strictly stationary.

REMARK 1.4.8

Norbert Wiener conjectured that **every** stationary sequence is a Bernoulli shift, which is not true. The truth is, almost every one is.

EXERCISE 1.4.9

Suppose $\{W_t\}_{t\in\mathbf{Z}}$ is a strong white noise. The **two-sided random walk** is defined as

$$X_t=\sum_{i=0}^{t}W_i+\sum_{i=t}^{-1}W_i$$

Show that $\{X_t\}_{t\in\mathbf{Z}}$ is first order stationary, but $\{X_t\}_{t\in\mathbf{Z}}$ is not second order stationary.

1.5 Weak versus Strong Stationary

Sadly, $\{X_t\}_{t\in\mathbf{Z}}$ is strictly stationary does not imply $\{X_t\}_{t\in\mathbf{Z}}$ is weakly stationary.

EXAMPLE 1.5.1

Suppose $X_t \overset{\text{iid}}{\sim}$ Cauchy Random Variables; that is,

$$\mathbb{P}(X_t \leq s) = \int_{-\infty}^{s} \frac{1}{\pi(1+x^2)}\, dx$$

Then, $\mathbb{E}[X_t]$ does not exist, and hence $\{X_t\}_{t\in\mathbf{Z}}$ cannot be weakly stationary. However, $\{X_t\}_{t\in\mathbf{Z}}$ is strictly stationary in this case since $\{X_t\}_{t\in\mathbf{Z}}$ is a strong white noise.

THEOREM 1.5.2

If $\{X_t\}_{t\in\mathbf{Z}}$ is strongly stationary and $\mathbb{E}[X_0^2] < \infty$, then $\{X_t\}_{t\in\mathbf{Z}}$ is weakly stationary.

Proof of: Theorem 1.5.2

Note that if $\{X_t\}_{t\in\mathbf{Z}}$ is strictly stationary,

$$(X_t) \overset{\text{d}}{=} (X_0)$$

so that $\mathbb{E}[X_t] = \mathbb{E}[X_0]$ which does not depend on t, and also

$$\mathbb{V}(X_t) = \mathbb{V}(X_0)$$

By the Cauchy-Schwarz inequality,

$$\gamma(t,s) = \mathsf{Cov}(X_t, X_s) \leq \mathbb{V}(X_t) < \infty$$

and suppose $t < s$,

$$\mathsf{Cov}(X_t, X_s) = \mathsf{Cov}(X_0, X_{s-t}) = f(|s-t|)$$

since it is shift-invariant, and hence if we shift everything over by t,

$$(X_t, X_s) \overset{\text{d}}{=} (X_{t-t}, X_{s-t}) \overset{\text{d}}{=} (X_0, X_{s-t})$$

DEFINITION 1.5.3: Gaussian process

$\{X_t\}_{t\in\mathbf{Z}}$ is said to be a **Gaussian process** (**Gaussian time series**) if for each $k \geq 1$, $i_1 < i_2 < \cdots < i_k$ we have

$$(X_{i_1}, \ldots, X_{i_k}) \sim \mathsf{MVN}(\boldsymbol{\mu}_k(i_1, \ldots, i_k), \boldsymbol{\Sigma}_{k\times k}(i_1, \ldots, i_k))$$

$$\boldsymbol{\mu}_k = \begin{bmatrix} \mathbb{E}[X_{i_1}] \\ \vdots \\ \mathbb{E}[X_{i_k}] \end{bmatrix} \qquad \boldsymbol{\Sigma}_{k\times k} = \mathsf{Cov}(X_{i_j}, X_{i_r})_{1\leq j,\, r\leq k}$$

THEOREM 1.5.4

If $\{X_t\}_{t\in\mathbf{Z}}$ is weakly stationary and a Gaussian process, then $\{X_t\}_{t\in\mathbf{Z}}$ is strictly stationary.

Proof of: Theorem 1.5.4

If $\{X_t\}_{t\in\mathbf{Z}}$ is weakly stationary, then $\mathbb{E}[X_t] = \mu$ for all t.

$$(X_{i_1}, \dots, X_{i_k}) \to \begin{bmatrix} \mathbb{E}[X_{i_1}] \\ \vdots \\ \mathbb{E}[X_{i_k}] \end{bmatrix} = \begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix} = \boldsymbol{\mu} = \begin{bmatrix} \mathbb{E}[X_{i_1+h}] \\ \vdots \\ \mathbb{E}[X_{i_k+h}] \end{bmatrix}$$

Also,

$$\begin{aligned} \mathbb{V}(X_{i_1}, \dots, X_{i_k}) &= \mathsf{Cov}(X_{i_j}, X_{i_r})_{1\le j,\, r\le k} \\ &= \mathsf{Cov}(X_0, X_{i_r - i_j})_{1\le j,\, r\le k} \\ &= \mathsf{Cov}(X_0, X_{i_r+h}, X_{i_r+h-(i_j+h)}) \\ &= \mathsf{Cov}(X_{i_j+h}, X_{i_r+h}) \\ &= \mathbb{V}(X_{i_1+h}, \dots, X_{i_k+h}) \end{aligned}$$

EXAMPLE 1.5.5

Using the Gaussian assumption

$$(X_{i_1}, \dots, X_{i_k}) \overset{\mathrm{d}}{=} \mathsf{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{k\times k}) \overset{\mathrm{d}}{=} (X_{i_1+h}, \dots, X_{i_k+h})$$

Hence $\{X_t\}_{t\in\mathbf{Z}}$ is strictly stationary in this case.

EXERCISE 1.5.6

Prove that if $\{X_t\}_{t\in\mathbf{Z}}$ is <u>not</u> weakly stationary; that is, either $\mathbb{E}[X_t]$ depends on t or $\gamma(t,s)$ does not depend on the lag, then $\{X_t\}_{t\in\mathbf{Z}}$ is <u>not</u> strictly stationary.

1.6 † Theoretical L2 Framework for Time Series

- $X_t = \lim_{h\to\infty} X_{h,t}$. In what sense does this limit exist?
- How "close" are two random variables X and Y?
- Is there a random variable that achieves

$$\inf_{y\in S} d(Y, S)$$

DEFINITION 1.6.1: L^2 space

Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The space L^2 is the set of random variables $X : \Omega \to \mathbf{R}$ measurable so that $\mathbb{E}[X^2] < \infty$.

DEFINITION 1.6.2: L^2-time series

We say that $\{X_t\}_{t\in\mathbf{Z}}$ is and L^2-time series if $X_t \in L^2$ for all $t \in \mathbf{Z}$.

L^2 is a Hilbert space when equipped with inner product, $X, Y \in L^2$.

$$\langle X, Y\rangle = \mathbb{E}[XY]$$

$\langle \cdot, \cdot \rangle$ is an inner product since it is

(1) Linear: $\langle aX + bY, Z\rangle = a\langle X, Z\rangle + b\langle Y, Z\rangle$.

(2) "Almost" Positive Definite: $\langle X, X\rangle = \mathbb{E}[X^2] = 0 \iff X = 0$ almost surely. Which implies $\mathbb{P}(X = 0) = 1$.

(3) Symmetric: $\langle X, Y\rangle = \langle Y, X\rangle$.

L^2 is complete with this inner product; that is, whenever $X_n \in L^2$ so that $\mathbb{E}[(X_n - X_m)^2] \to 0$ as $n, m \to \infty$, then there exists $X \in L^2$ so that $X_n \to X$; that is, $\mathbb{E}[(X_n - X)^2] \to 0$. This follows from the "famous" Riesz-Fischer Theorem.

1.6.1 Useful Tools for Time Series

(1) **Existence of Limits**

$$X_{t,n} = \sum_{j=0}^{n} \psi_j \varepsilon_{t-j}$$

$\{\varepsilon_t\}_{t\in\mathbf{Z}}$ is a strong white noise. Since for $n > m$,

$$\mathbb{E}[(X_{t,n} - X_{t,m})^2] = \mathbb{E}\left[\left(\sum_{j=m+1}^{n} \psi_j \varepsilon_{t-j}\right)^2\right] = \sum_{j=m+1}^{n} \psi_j^2 \sigma_\varepsilon^2 \to 0 \text{ as } n, m \to \infty$$

only if $\sum_{j=0}^{\infty} \psi_j^2 < \infty$, then there **must** exist a random variable X_t (by the completeness of L^2), so that

$$X_t = \lim_{n\to\infty} X_{t,n} = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$$

(2) **Projection Theorem and Forecasting**. Forecasting can be often cast as finding a random variable Y among a collection of possible forecasts $\mathcal{M}$ (e.g. $\mathcal{M} = \text{Span}(X_T, \dots, X_1)$) so that

$$Y = \arg\inf_{Z\in\mathcal{M}} \mathbb{E}[(X_{T+h} - Z)^2]$$

When $\mathcal{M}$ is a closed linear subspace of L^2, the Projection Theorem guarantees that such a Y exists, and it must satisfy

$$\langle X_{T+h} - Y, Z\rangle = 0 \quad \forall Z \in \mathcal{M}$$

must be in the orthogonal complement.

1.7 Signal and Noise Models

"Ideally," a time series that we are considering was generated from a stationary process. If so, we can pool data to estimate the processes underlying structure (e.g. its marginal distribution and serial dependence structure)

Most time series are evidently not stationary.

Looking back at Figure 1.1:

- Mean appears to increase, so it is not first order stationary;
- Variability also appears to increase, so it is not second order stationary;
- Therefore, it is not strictly stationary.

Signal and Noise Model: $X_t = S_t + \varepsilon_t$

- S_t is the **deterministic** "signal" or "trend" of the series

- ε_t is the "noise" added to the signal satisfying $\mathbb{E}[\varepsilon_t] = 0$, hence $\mathbb{E}[X_t] = \mathbb{E}[S_t + \varepsilon_t] = \mathbb{E}[S_t]$. There exists a (strong) white noise $\{W_t\}_{t\in\mathbf{Z}}$ so that

$$\varepsilon_t = g(W_t, W_{t-1}, ...) \quad \text{[Stationary Noise]}$$

$$\varepsilon_t = g_t(W_t, W_{t-1}, ...) \quad \text{[Non-stationary Noise]}$$

The terms $\{W_t\}_{t\in\mathbf{Z}}$ are often called the "innovations" or "shocks" during the random behaviour of X_t.

g is used to try to capture noise that can potentially have serial dependence.

EXAMPLE 1.7.1

An example of a function g so that $\varepsilon_t = g_t(W_t, W_{t-1}, ...)$ might be a **random walk**; that is, $\varepsilon_t = \sum_{j=0}^{t} W_j$. Another example could be the **changing variance models**; that is, $\varepsilon_t = \sigma(t)W_t$.

Our goal is to estimate S_t, and then infer the structure of ε_t.

In Figure 1.2, the model appears to be non-stationary (trending upwards over time), so we might try the signal and noise model. We might posit a linear trend, or even higher order functions.

For the temperature data, we may posit that

$$S_t = \beta_0 + \beta_1 t \quad \text{[Linear Trend]}$$

The trend may be estimated by ordinary least squares (OLS). We choose β_0 and β_1 to minimize

$$\sum_{t=1}^{T} [X_t - (\beta_0 + \beta_1 t)]^2$$

This can be done in R using the `lm()` command, and can easily be computed with calculus. Figure 1.4 is a small example of the global temperature data superimposed with `lm()`'s estimate.

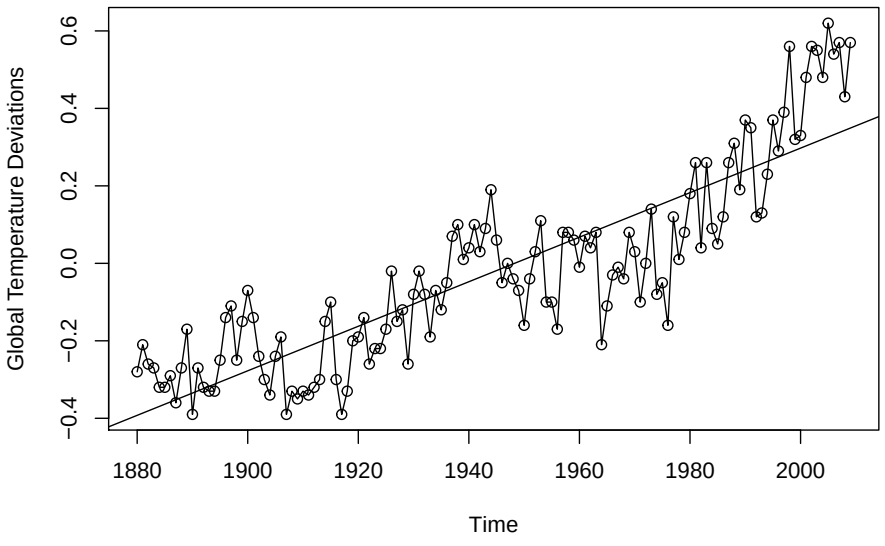

Figure 1.4: OLS estimate of linear trend

```
# Figure 1.4
fit <- lm(gtemp ~ time(gtemp), na.action = NULL)
plot.ts(gtemp, type = "o", ylab = "Global Temperature Deviations")
abline(fit)
```

Let's introduce some terminology about trends.

DEFINITION 1.7.2: Detrended time series

Detrending a time series constitutes computing the residuals based on an estimate for the signal/trend. A **detrended time series** is a time series of such residuals.

1. Estimate $S_t \to \hat{S}_t$
2. Detrend series: $X_t - \hat{S}_t = Y_t$ where Y_t is the "detrended" series.

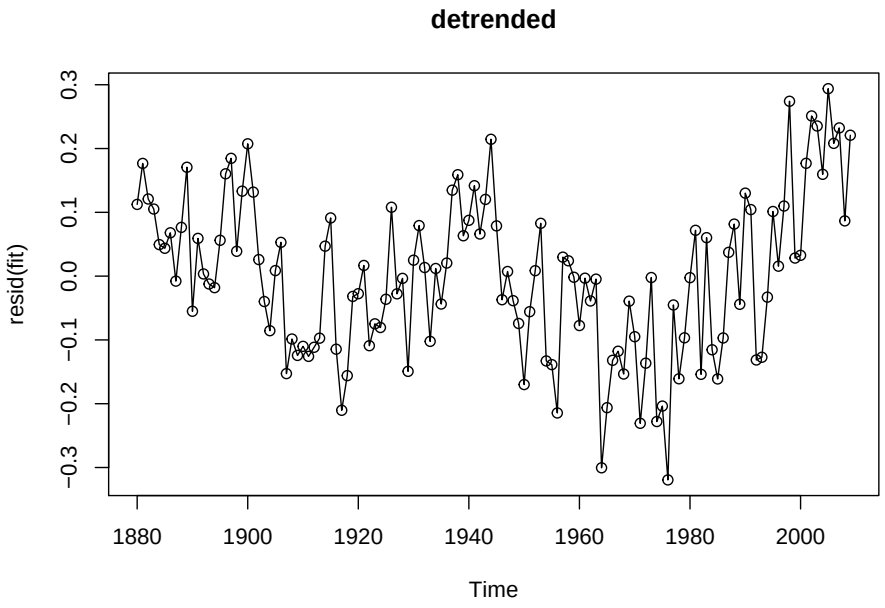

Figure 1.5: Residuals of OLS fit.

```
# Figure 1.5
plot(resid(fit), type = "o", main = "detrended")
```

In Figure 1.5: If trend is now zero, there appears to be a substantial serial dependence remaining in the time series.

1.8 Time Series Differencing

Signal and Noise Model: $X_t = S_t + \varepsilon_t$. Hopefully, upon estimating S_t with $\hat{S}_t$, we find $X_t - \hat{S}_t = \hat{\varepsilon}_t$ (detrended series) looks reasonably stationary. If the residuals would be reasonably stationary, we might proceed in estimating their underlying structure of $\{\hat{\varepsilon}_t\}_{t=1,\ldots,T}$. as if it were stationary. In particular, we might try to estimate their marginal distributions and/or their serial dependence structure. If we thought those estimates were reasonably good, we would have a good idea of how the time series X_t behaves.

Random Walk with Drift Model. Let ε_t be a strong white noise.

$$\begin{aligned} X_t &= \delta + X_{t-1} + \varepsilon_t \\ &= \delta + \delta + X_{t-2} + \varepsilon_{t-1} + \varepsilon_t \\ &= \delta + \delta + \delta + X_{t-3} + \varepsilon_{t-2} + \varepsilon_{t-1} + \varepsilon_t \\ &\vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad t \text{ times} \\ &= t\delta + X_0 + \sum_{j=1}^{t} \varepsilon_j \end{aligned}$$

where we note that $t\delta + X_0 = S_t$ is a linear signal, and $\sum_{j=1}^{t} \varepsilon_j$ is a random walk noise.

Notice that under the Random Walk Model.

$$X_t - X_{t-1} = \nabla X_t = \delta + \varepsilon_t$$

So, if X_t follows a random walk model, the series $Y_t = \delta X_t$ should behave like a white noise shifted by δ.

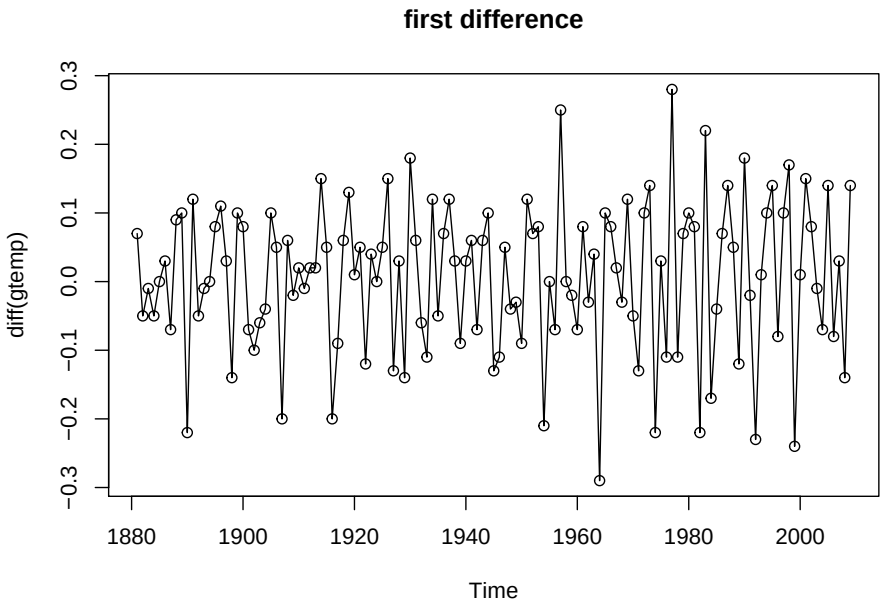

Figure 1.6: First differenced series. Average of first differenced series is $\hat{\delta} \approx 0.0066$

```
# Figure 1.6
plot(diff(gtemp), type = "o", main = "first difference")
```

In Figure 1.6: To see what this looks like in this temperature example, here is a plot of $\nabla X_t = X_t - X_{t-1}$ for the temperature deviation data. As you can see if you look at this compared to the detrended series using linear trend, I would say this series looks much more like a white noise (there does not appear to be any discernible patterns in this first difference). If you calculate the mean of this first difference series, that would be an estimator for the drift term in the random walk model which here is ≈ 0.0066.

DEFINITION 1.8.1: Differenced time series

Differencing a time series constitutes computing the difference between successive terms.
A **differenced time series** is a time series of such differences. The first differenced series is denoted

$$\nabla X_t = X_t - X_{t-1}$$

and is the series of length $T-1$, namely

$$X_2 - X_1, X_3 - X_2, \dots, X_T - X_{T-1}$$

Higher order differences are calculated recursively, so

$$\nabla^d X_t = \nabla^{d-1} \nabla X_t$$

where ∇^d is the d^{th} order difference and we define $\nabla^0 X_t = X_t$.

Detrending and Differencing are both ways of reducing a (potentially non-stationary) time series to an approximately stationary series.

Differencing vs. Detrending

Pros:

- Differencing does not require the parameter estimation (don't need to estimate S_t).
- Higher order differencing can reduce even very "trendy" series to look more like noise.

Cons:

- Differencing can "wash away" features of the series, and introduce more complicated structures.
- The trend is often of interest, and good estimates of the trend lead to improved long-range forecasts.

EXAMPLE 1.8.2: Differencing can complicate time series

$X_t = W_t$ where W_t is a strong white noise.

$$\nabla X_t = W_t - W_{t-1} = Y_t$$

$$\gamma_X(h) = \mathsf{Cov}(X_t, X_{t+h}) = \begin{cases} \sigma_W^2 & h = 0 \\ 0 & h \geq 1 \end{cases}$$

More complicated:

$$\gamma_Y(h) = \mathsf{Cov}(Y_t, Y_{t+h}) = \begin{cases} 2\sigma_W^2 & h = 0 \\ -\sigma_W^2 & h = 1 \\ 0 & h \geq 2 \end{cases}$$

```
# Figure 1.7
par(mfrow = c(2, 1))
plot(diff(gtemp), main = "first difference Temp data")
plot(rnorm(gtemp),
     type = "l",
     main = "white noise",
     ylab = "w")
```

In Figure 1.7: If these two series behave in the same way, then it stands to reason that

$$g(\varepsilon_t, \varepsilon_{t-1}, \dots) = \varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_{\text{temp}})$$

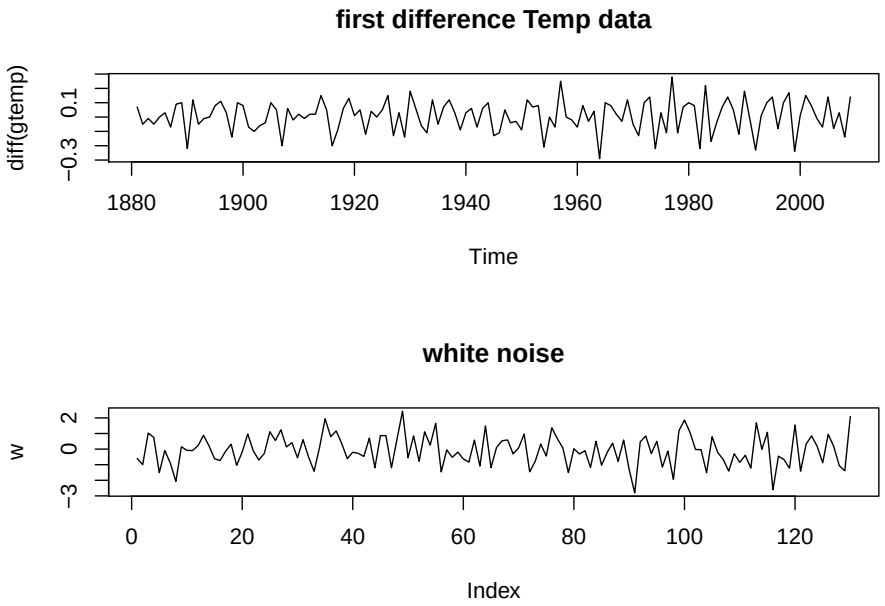

Figure 1.7: First Difference and White Noise

Chapter 2

Time Series Regression

2.1 Autocorrelation and Empirical Autocorrelation

Usually through either detrending or differencing, we arrive at a series $\{X_t\}_{t\in\mathbf{Z}}$ that we may consider as stationary.

Given such a series, we wish to estimate a function g, so that

$$X_t = g(W_t, W_{t-1}, ...)$$

$\{W_t\}_{t\in\mathbf{Z}}$ is a "innovation" sequence (strong white noise) which could admit serial dependence, etc.

In a first pass, it's reasonable to assume that g is a linear function.

DEFINITION 2.1.1: Linear process

A time series $\{X_t\}_{t\in\mathbf{Z}}$ is said to be a **linear process** if there exists a strong white noise $\{W_t\}_{t\in\mathbf{Z}}$ and coefficient $\{\psi_\ell\}_{\ell\in\mathbf{Z}}$ where $\psi_\ell \in \mathbf{R}$, so that

$$\sum_{\ell=-\infty}^{\infty} |\psi_\ell| < \infty$$

and

$$X_t = \sum_{\ell=-\infty}^{\infty} \psi_\ell W_{t-\ell}$$

Note that the sum defining X_t is well-defined as a limit in L^2. Also, we must require that $\mathbb{V}(W_{t-\ell}) < \infty$.

DEFINITION 2.1.2: Causal linear process

We say $\{X_t\}_{t\in\mathbf{Z}}$ is a **causal linear process** if

$$X_t = \sum_{\ell=0}^{\infty} \psi_\ell W_{t-\ell}$$

Note that X_t only depends on W's in the "past."

EXAMPLE 2.1.3

$X_t = W_t$ is a linear process, so all ψ's are 0, except for $\psi_0 = 1$ which is a strong white noise sequence.

REMARK 2.1.4

Linear processes are **strictly stationary** since they can be written as Bernoulli-shifts.

EXAMPLE 2.1.5

$X_t = W_t + \theta W_{t-1}$ where $\{W_t\}_{t\in\mathbf{Z}}$ is a strong white noise with finite variance, and X_t is a linear process.

$$\gamma_X = \begin{cases} (1+\theta^2)\sigma_W^2 & h = 0 \text{always non-zero} \\ \theta\sigma_W^2 & h = 1 \\ 0 & h \geq 2 \end{cases}$$

$\gamma_X(h)$ non-zero for $h \geq 1$ only where "lagged" terms in the linear process are non-zero. Suggests a way of sleuthing out what

$$g(W_t, W_{t-1}, ...) = \sum_{\ell=0}^{\infty} \psi_\ell W_{t-\ell}$$

must look like.

DEFINITION 2.1.6: Autocorrelation function

Suppose $\{X_t\}_{t\in\mathbf{Z}}$ is weakly stationary. The **autocorrelation function** (ACF) of $\{X_t\}_{t\in\mathbf{Z}}$ is

$$\rho_X(h) = \frac{\gamma(h)}{\gamma(0)} \quad (h \geq 0)$$

Note since $\gamma(0) = \mathbb{V}(X_t) = \mathbb{V}(X_0)$ (since the process is stationary),

$$|\gamma(h)| = |\mathsf{Cov}(X_t, X_{t+h})| \underset{\text{CS}}{\leq} \sqrt{\underbrace{\mathbb{V}(X_t)\mathbb{V}(X_{t+h})}_{\text{Same \# by stationarity}}} = \mathbb{V}(X_0)$$

Hence, $|\rho(h)| \leq 1 \implies -1 \leq \rho(h) \leq 1$.

Estimating $\gamma(h)$ and $\rho(h)$:

$$\gamma(h) = \mathsf{Cov}(X_t, X_{t+h}) = \mathbb{E}[(X_t - \mu)(X_{t+h} - \mu)]$$

where $\mu = \mathbb{E}[X_t]$. Hence, a sensible estimator is

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} X_t = \bar{X}$$

which is the **sample mean** (**time series average**).

$$\hat{\gamma}(h) = \frac{1}{T}\sum_{t=1}^{T-h}(X_t - \bar{X})(X_{t+h} - \bar{X}) \approx \frac{1}{T-h}\sum_{t=1}^{T-h}(X_t - \bar{X})(X_{t+h} - \bar{X})$$

where $(X_t - \bar{X})(X_{t+h} - \bar{X})$ is the averaging over centred terms h-time steps apart.

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

EXAMPLE 2.1.7

$X_t = W_t$ where $\{W_t\}_{t\in\mathbf{Z}}$ is a strong white noise with $\mathbb{V}(W_t) = \sigma_W^2 < \infty$.

$$\gamma_X(h) = \begin{cases} \sigma_W^2 & h = 0 \\ 0 & h \geq 1 \end{cases}$$

Therefore,

$$\rho_X(h) = \begin{cases} 1 & h = 0 \\ 0 & h \geq 1 \end{cases}$$

Note that it's always the case that

$$\rho(0) = \frac{\gamma(0)}{\gamma(0)} = 1$$

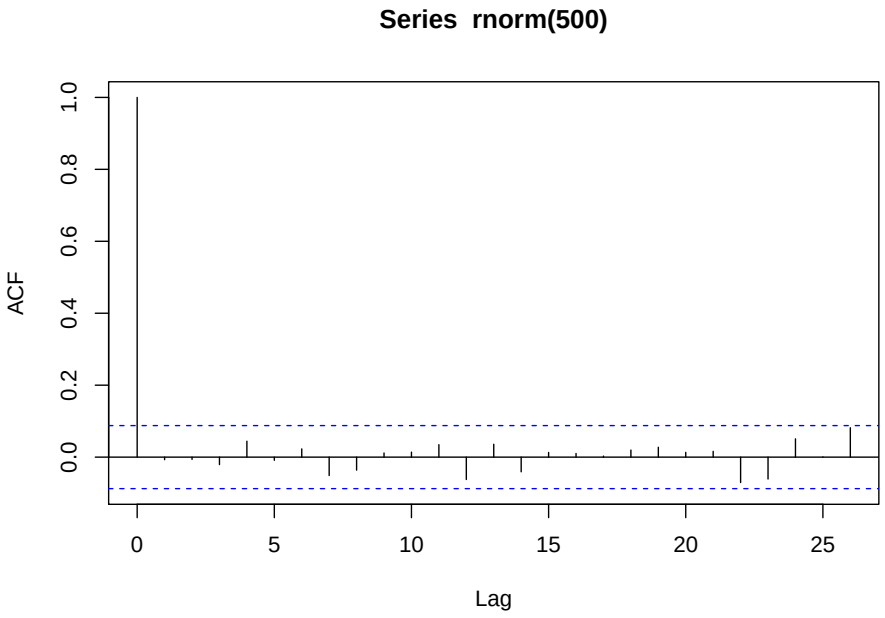

Figure 2.1: Quarterly Johnson and Johnson Earnings

```
# Figure 2.1
acf(rnorm(500))
```

In Figure 2.1: Let's then have a look at what the empirical autocorrelation function looks like when we apply it to a strong white noise sample. In this case, we are considering a strong Gaussian white noise with variance 1. This is what the sample ACF looks like. What we're plotting here is on the x-axis we have the lags h, and on the y-axis we have the magnitudes of the autocorrelation $\hat{\rho}(h)$. What we're seeing here is $\hat{\rho}(0) = 1$ (by definition). However, for lags other than zero, for the other autocorrelations plotted, we can see that they are relatively small compared to $\hat{\rho}(0) = 1$, which is the point of the blue lines (explained in the next lecture). The basic interpretation of blue lines is that if an autocorrelation would go inside the blue lines then you could imagine that it would be consistent with the series being a strong white noise, which is what we observe here. There's small violations that can occur by sheer chance.