

Lecture 12

Time Series Analysis I

DSA 8020 Statistical Methods II
April 4-8, 2021

Time Series Data

Time Series Models

Objectives of Time
Series Analysis

A Case Study

Whitney Huang
Clemson University

Agenda

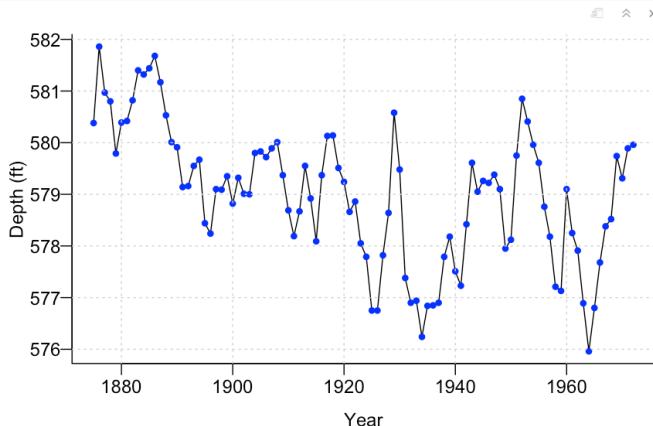
- 1 Time Series Data
- 2 Time Series Models
- 3 Objectives of Time Series Analysis
- 4 A Case Study

Level of Lake Huron 1875–1972

Annual measurements of the level of Lake Huron in feet.

[Source: [Brockwell & Davis, 1991](#)]

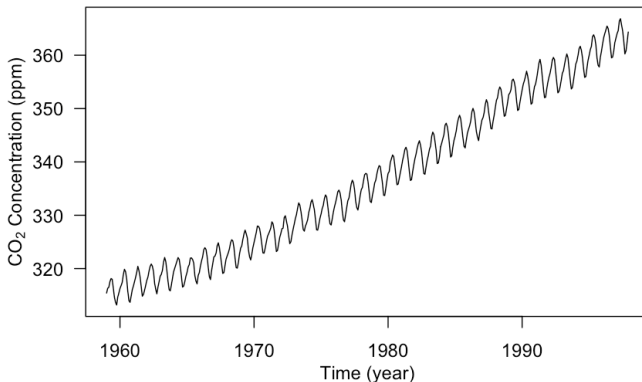
```
## {r}
par(mar = c(3.2, 3.2, 0.5, 0.5), mgp = c(2, 0.5, 0), bty = "L")
data(LakeHuron)
plot(LakeHuron, ylab = "Depth (ft)", xlab = "Year", las = 1)
points(LakeHuron, cex = 0.8, col = "blue", pch = 16)
grid()
```



Mauna Loa Atmospheric CO₂ Concentration

Monthly atmospheric concentrations of CO₂ at the Mauna Loa Observatory [Source: [Keeling & Whorf, Scripps Institution of Oceanography \(SIO\)](#)]

```
```{r}
data(co2)
par(mar = c(3.8, 4, 0.8, 0.6))
plot(co2, las = 1, xlab = "", ylab = "")
mtext("Time (year)", side = 1, line = 2)
mtext(expression(paste("CO"[2], " Concentration (ppm)")), side = 2, line = 2.5)
```
```



US Unemployment Rate 1948 Jan. – 2021 July

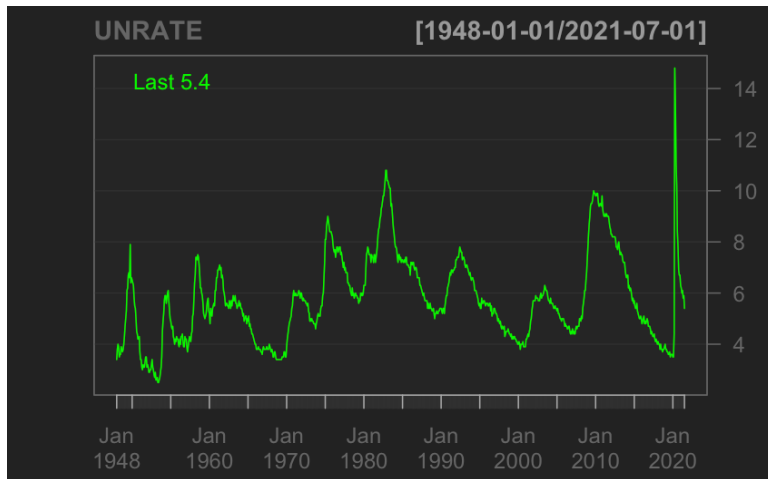
[Source: St. Louis Federal Reserve Bank's FRED system]

Time Series Data

Time Series Models

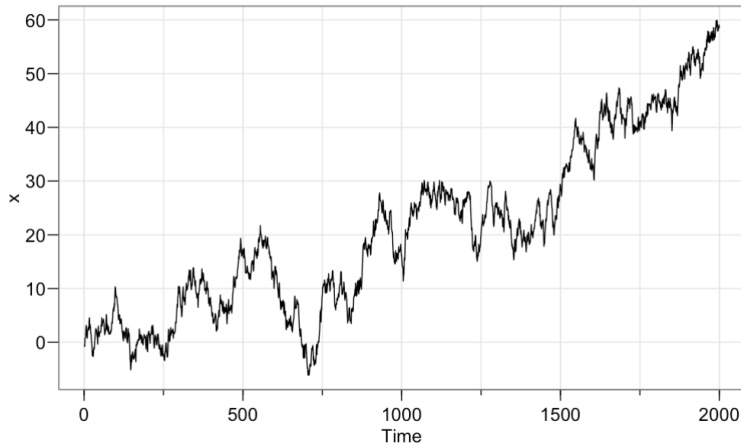
Objectives of Time
Series Analysis

A Case Study




A Simulated Time Series

```
## {r}  
set.seed(123)  
w <- rnorm(2000); x <- cumsum(w); tsplot(x, las = 1)  
##
```



- A **time series** is a set of observations $\{y_t, t \in T\}$ made sequentially in time (t) with the index set T
 - $T = \{0, 1, 2, \dots, T\} \subset \mathbb{Z} \Rightarrow$ **discrete-time time series**
 - $T = [0, T] \subset \mathbb{R} \Rightarrow$ **continuous-time time series**
- A discrete-time time series might be intrinsically discrete or might arise from a underlying continuous-time time series via
 - sampling (e.g., instantaneous wind speed)
 - aggregation (e.g., daily accumulated precipitation amount)
 - extrema (e.g., daily maximum temperature)
- We will focus on dealing with **discrete-time real-valued** ($Y_t \in \mathbb{R}$) **time series**

- Start with a **time series plot**, i.e., to plot y_t versus t 
- Look at the following:
 - Are there abrupt changes?
 - Are there “outliers”?
 - Is there a need to transform the data?
- Examine the **trend**, **seasonal components**, and the “noise” term

● Trends

- One can think of trend, μ_t , as continuous changes, usually in the mean, over longer time scales \Rightarrow *“the essential idea of trend is that it shall be smooth”* - [Kendall, 1973]
- Usually the form of the trend is unknown and needs to be estimated. When the trend is removed, we obtain a **detrended** series

● Seasonal or periodic components

- A seasonal component s_t constantly repeats itself in time, i.e., $s_t = s_{t+kd}$
- We need to estimate the form and/or the period d of the seasonal component to **deseasonalize** the series

● The “noise” process

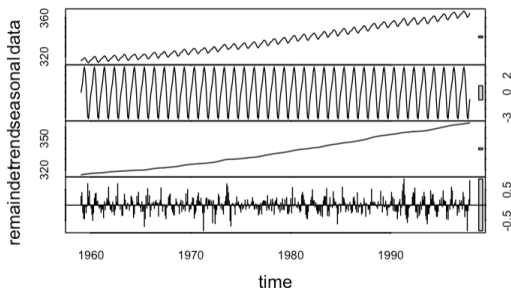
- The noise process, η_t , is the component that is neither trend nor seasonality
- We will focus on finding plausible (typically stationary) statistical models for this process

Combining Trend, Seasonality, and Noise Together

There are two commonly used approaches

- Additive model:

$$y_t = \mu_t + s_t + \eta_t, \quad t = 1, \dots, T$$



- Multiplicative model:

$$y_t = \mu_t s_t \eta_t, \quad t = 1, \dots, T$$

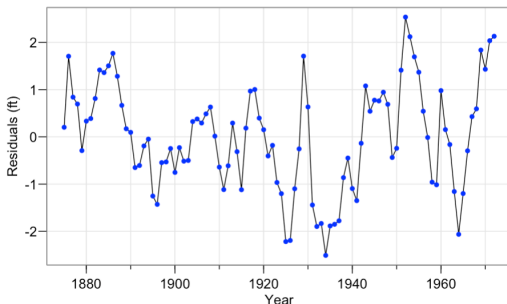
If all $\{y_t\}$ are positive then we obtain the additive model by taking logarithms:

$$\log y_t = \log \mu_t + \log s_t + \log \eta_t, \quad t = 1, \dots, T$$

Time Series Models

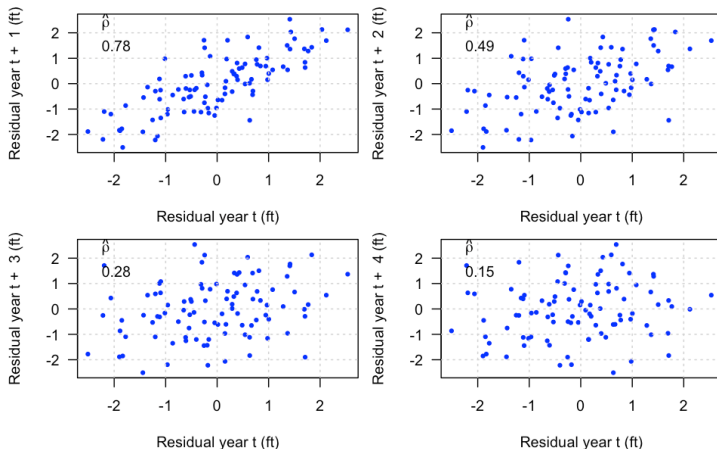
Lake Huron Time Series

- **Time series analysis** is the area of statistics which deals with the analysis of **dependency** between different observations (typically $\{\eta_t\}$)
- Some key features of the Lake Huron time series:
 - decreasing trend
 - some “random” fluctuations around the decreasing trend
- We extract the “noise” component by assuming a linear trend



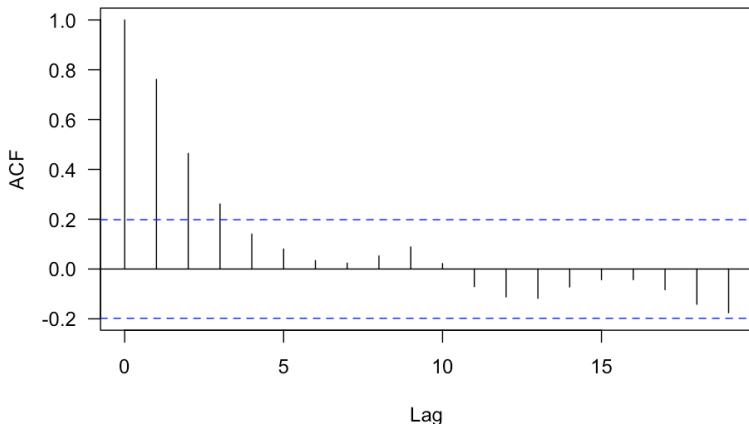
Exploring the Temporal Dependence Structure of $\{\eta_t\}$

$\{\eta_t\}$ exhibit some temporal dependence structure, that is, the nearby (in time) values tend to be more alike than those far apart values. To see this, let's make a few time lag plots



Further Exploration of the Temporal Dependence Structure

Let's plot the correlation as a function of the time lag



We will use this information to suggest an appropriate model

- A **time series model** is a probabilistic model for $\{Y_t : t \in T\}$ that describes ways that the series data $\{y_t\}$ could have been generated
- Will try to keep our models for $\{Y_t\}$ simple by assuming **stationarity** \Rightarrow characteristic of the distribution of $\{Y_t\}$ does not depend on the time points, only on the “time lag”
- While most time series are not stationary, one either remove or model the non-stationary parts (e.g., de-trend or de-seasonalization) so that we are only left with a stationary component $\{\eta_t\}$.

- The **mean function** of $\{\eta_t\}$ is

$$\mu_t = E[\eta_t], \quad t \in T$$

- The **autocovariance function** of $\{\eta_t\}$ is

$$\gamma(t, t') = \text{Cov}(\eta_t, \eta_{t'}) = E[(\eta_t - \mu_t)(\eta_{t'} - \mu_{t'})], \quad t, t' \in T,$$

when $t = t'$ we obtain $\gamma(t, t') = \text{Cov}(\eta_t, \eta_t) = \text{Var}(\eta_t) = \sigma_t^2$,
the variance function of Y_t

The autocorrelation function (ACF) of $\{\eta_t\}$ is

$$\rho(t, t') = \text{Corr}(\eta_t, \eta_{t'}) = \frac{\gamma(t, t')}{\sqrt{\gamma(t, t)\gamma(t', t')}}.$$

It measures the strength of linear association between Y_t and $Y_{t'}$

Properties:

- 1 $-1 \leq \rho(t, t') \leq 1, \quad t, t' \in T$
- 2 $\rho(t, t') = \rho(t', t), \quad \forall t, t' \in T; \rho(t, t) = 1, \quad \forall t \in T$
- 3 $\rho(t, t')$ is a non-negative definite function

Partial autocorrelation function (PACF) is a conditional correlation, i.e., the correlation at two time points given the information at all other time points

We will try to keep our **models for $\{\eta_t\}$** as simple as possible by assuming **stationarity**, meaning that characteristic of $\{\eta_t\}$ does not depend on the time points, only on the “time lag”:

- $E[\eta_t] = 0, \quad \forall t \in T$

- $\text{Cov}(\eta_t, \eta_{t'}) = \gamma(t' - t) = \text{Cov}(\eta_{t+s}, \eta_{t'+s})$

⇒ autocorrelation function (ACF):

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

Let $\{Z_t\}$ be independent and identical random variables that follow $N(0, \sigma^2)$

- Moving Average Processes (MA(q)):

$$\eta_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} \cdots + \theta_q Z_{t-q}$$

Let $\{Z_t\}$ be independent and identical random variables that follow $N(0, \sigma^2)$

- Moving Average Processes (MA(q)):

$$\eta_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} \cdots + \theta_q Z_{t-q}$$

- Autoregressive Processes (AR(p)):

$$\eta_t = \phi_1 \eta_{t-1} + \phi_2 \eta_{t-2} + \cdots + \phi_p \eta_{t-p} + Z_t$$

Let $\{Z_t\}$ be independent and identical random variables that follow $N(0, \sigma^2)$

- Moving Average Processes (MA(q)):

$$\eta_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} \cdots + \theta_q Z_{t-q}$$

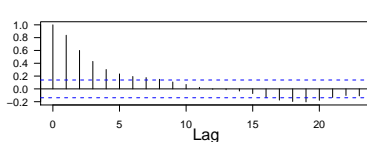
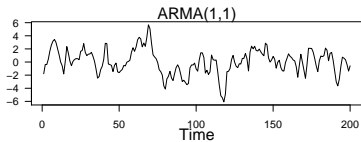
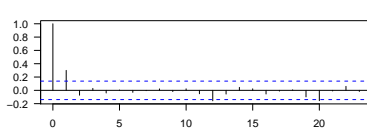
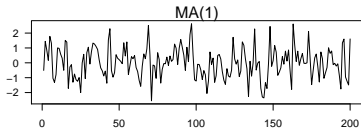
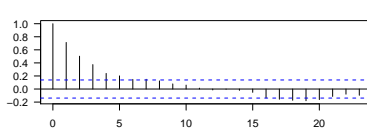
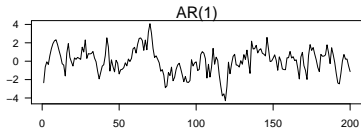
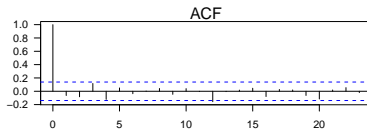
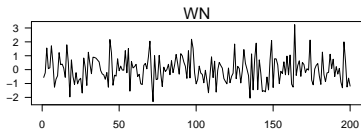
- Autoregressive Processes (AR(p)):

$$\eta_t = \phi_1 \eta_{t-1} + \phi_2 \eta_{t-2} + \cdots + \phi_p \eta_{t-p} + Z_t$$

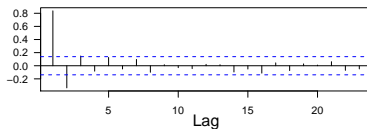
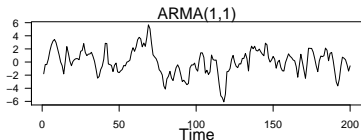
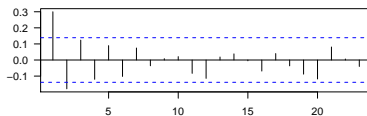
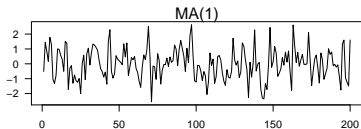
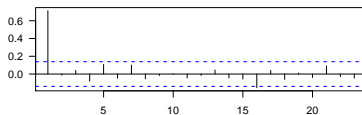
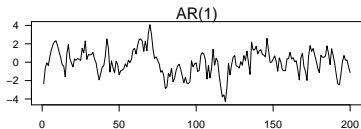
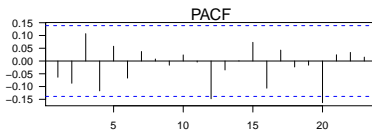
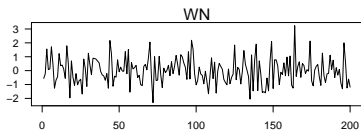
- Autoregressive Moving Average Processes ARMA(p,q):

$$\eta_t = \phi_1 \eta_{t-1} + \phi_2 \eta_{t-2} + \cdots + \phi_p \eta_{t-p} + Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \cdots + \theta_q Z_{t-q}$$

ACF Plots



PACF Plots



Identification of ARMA Models using ACF/PACF Plots

Use the ACF and PACF together to identify possible models. The following table gives some rough guidelines. Unfortunately, it's not a well-defined process and some guesswork is usually needed

| | ACF | PACF |
|--------------|------------------------|------------------------|
| $AR(p)$ | Tails off | Cuts off after lag p |
| $MA(q)$ | Cuts off after lag q | Tails off |
| $ARMA(p, q)$ | Tails off | Tails off |

We wish to test:

$H_0 : \{e_1, e_2, \dots, e_T\}$ is an i.i.d. noise sequence;

$H_1 : H_0$ is false,

where $\{e_t\}$ are the residuals after fitting a model to $\{\eta_t\}$

Test statistic:

$$Q_{LB} = T(T-2) \sum_{h=1}^k \frac{\hat{\rho}^2(h)}{T-h} \sim \chi_k^2.$$

Ljung-Box test can be carried out in R using the function

`Box.test`

Objectives of Time Series Analysis

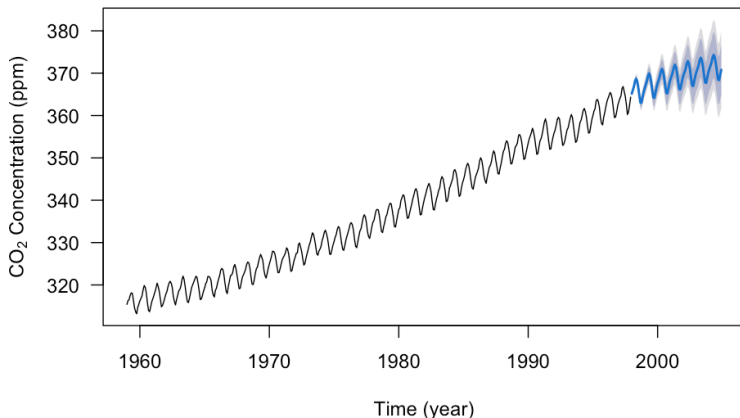
Some Objectives of Time Series Analysis

- **Modeling:** Find a **statistical model** that adequately explains the observed time series
- For example, identify a model which can account for the fact that the depths of Lake Huron are correlated with different years and with a decreasing long-term trend
- The fitted model can be used for further **statistical inference**, for instance, to answer the question like: **Is there evidence of decreasing trend in the Lake Huron depths?**

Some Objectives of Time Series Analysis, Cont'd

Forecasting is perhaps the most common objective. One observe a time series of given length and wish to **predict** or **forecast** future values of the time series based on those already observed.

Forecasts from TBATS(1, {3,1}, -, {<12,5>})



Some Objectives of Time Series Analysis, Cont'd

- **Adjustment:** an example would be [seasonal adjustment](#), where the seasonal component is estimated and then removed in order to better understand the underlying trend
- **Simulation:** use a time series model (which adequately describes a physical process) as a surrogate to *simulate repeatedly in order to approximate how the physical process behaves*
- **Control:** adjust various [input \(control\)](#) parameters so that the time series fits closer to a given standard (many examples from statistical quality control)

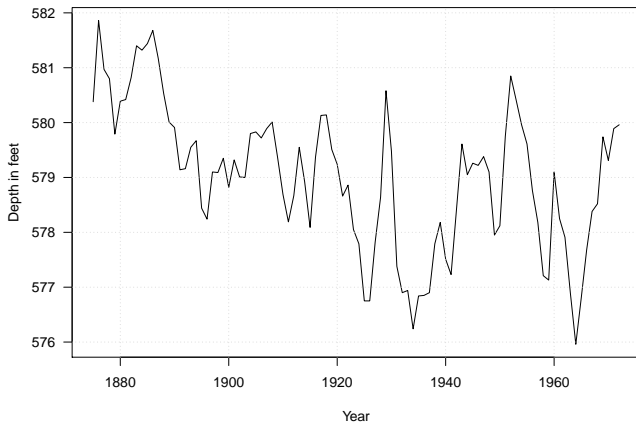
Lake Huron Case Study



Source: <https://www.worldatlas.com/articles/what-states-border-lake-huron.html>

- Detrending
- Model fitting and selection
- Forecasting

Annual Measurements of the Level of Lake Huron

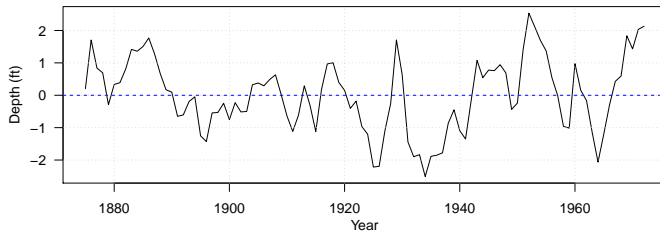
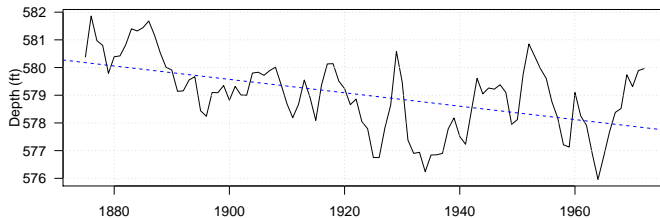


There seems to be a decreasing trend \Rightarrow need to estimate the trend to get the detrended series

Plots of the Trend and Residuals

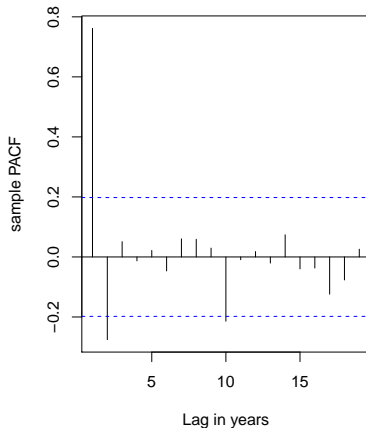
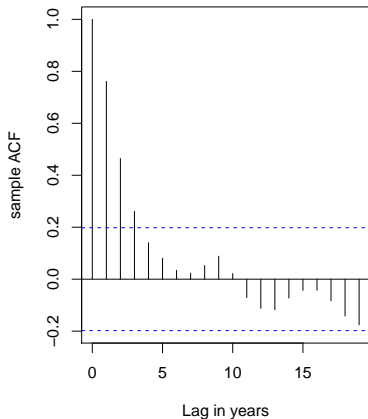
$$y_t = \underbrace{\mu_t}_{\text{trend}} + \underbrace{\eta_t}_{\text{residual}}$$

where we **assume** $\mu_t = \alpha + \beta t$, i.e., a **linear trend in time**

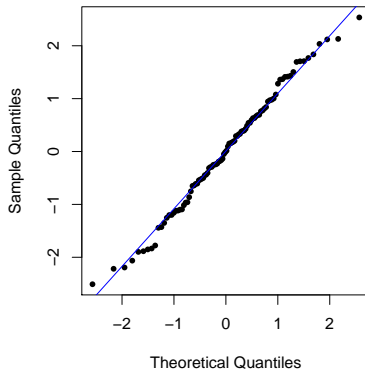
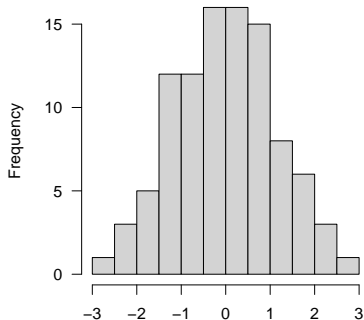


ACF and PACF Plots

- Tapering pattern in ACF \Rightarrow need to include AR terms
- Significant PACF values at the first 2 lags \Rightarrow a AR(2) may be appropriate



Assessing Normality Assumption for η_t



Fitting AR(2)

```
> (ar2.model <- arima(deTrend, order = c(2, 0, 0)))
```

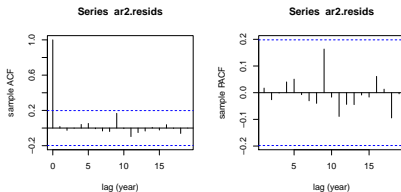
Call:

```
arima(x = deTrend, order = c(2, 0, 0))
```

Coefficients:

| | ar1 | ar2 | intercept |
|------|--------|---------|-----------|
| | 1.0047 | -0.2919 | 0.0196 |
| s.e. | 0.0977 | 0.1004 | 0.2351 |

sigma^2 estimated as 0.4571: log likelihood = -101.25, aic = 210.5



```
> Box.test(ar2.resids, type = "Ljung-Box")
```

Box-Ljung test

data: ar2.resids

X-squared = 0.029966, df = 1, p-value = 0.8626

We can conduct model selection by using, for example, AIC

```
> ar1.model <- arima(deTrend, order = c(1, 0, 0))  
> ar2.model <- arima(deTrend, order = c(2, 0, 0))  
> arma21.model <- arima(deTrend, order = c(2, 0, 1))  
> AIC(ar1.model); AIC(ar2.model); AIC(arma21.model)  
[1] 216.5835  
[1] 210.5032  
[1] 212.1784
```

Fitting AR(2) + a Linear Trend

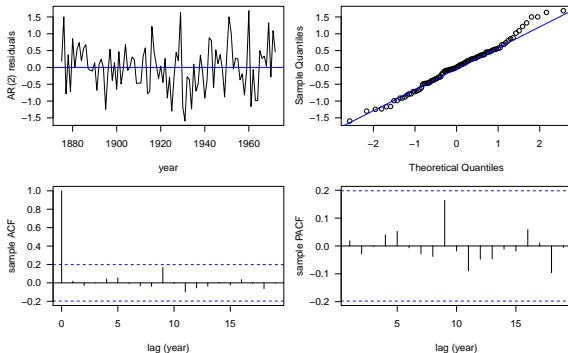
```
> library(forecast)
> (fit <- Arima(LakeHuron, order = c(2, 0, 0), include.drift = T))
Series: LakeHuron
ARIMA(2,0,0) with drift
```

Coefficients:

| | ar1 | ar2 | intercept | drift |
|------|--------|---------|-----------|---------|
| | 1.0048 | -0.2913 | 580.0915 | -0.0216 |
| s.e. | 0.0976 | 0.1004 | 0.4636 | 0.0081 |

sigma² estimated as 0.476: log likelihood=-101.2

AIC=212.4 AICc=213.05 BIC=225.32



10-Year-Ahead Forecasts

