Lecture 12
Time Series Analysis I

DSA 8020 Statistical Methods II

Whitney Huang
Clemson University

Agenda

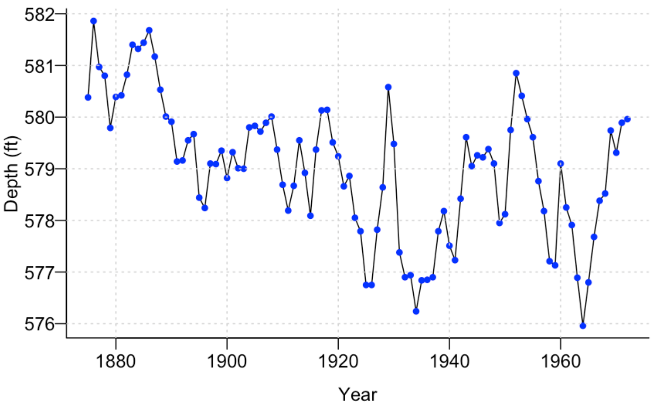
1. **Background**
2. **Time Series Models**
3. **A Case Study**

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()
```

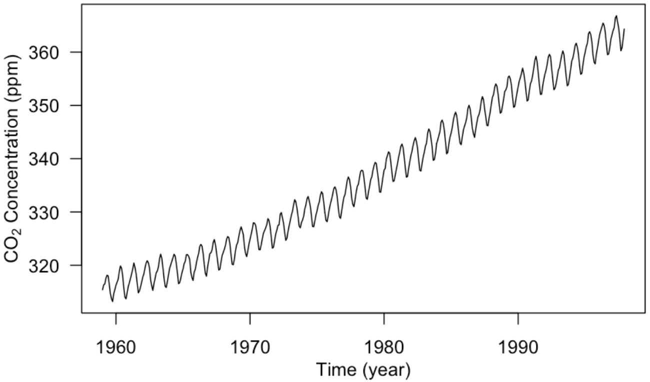


Return

Mauna Loa Atmospheric CO_2 Concentration

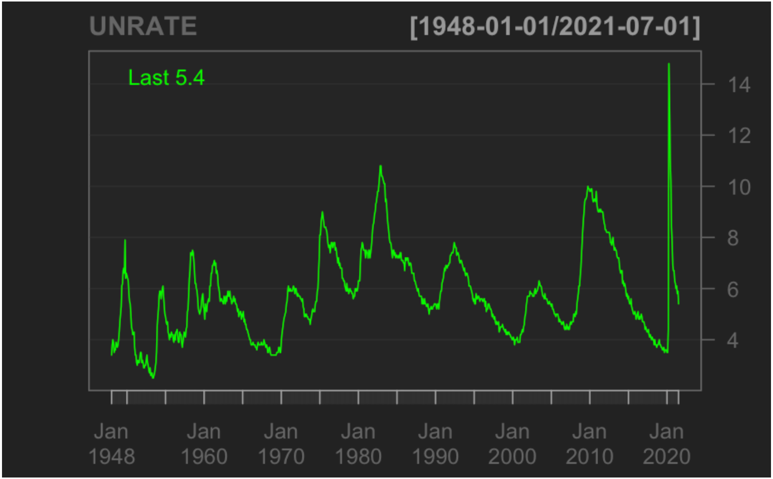
Monthly atmospheric concentrations of CO_2 at the Mauna Loa Observatory [Source: Keeling & Whorf, Scripps Institution of Oceanography (SIO)]

```{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)
```



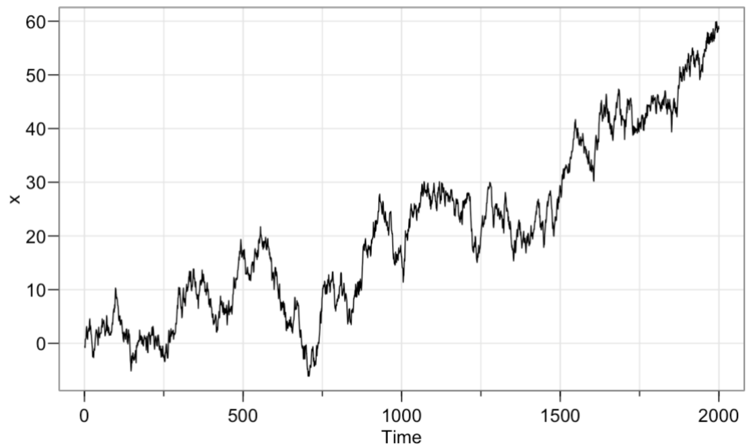
US Unemployment Rate 1948 Jan. – 2021 July

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



A Simulated Time Series

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



Time Series Data

- A time series is a set of observations $\{y_t, t \in T\}$ made sequentially in time (t) with the index set T
  - $T = \{0, 1, 2, \cdots, T\} \subset \mathbb{Z} \Rightarrow$ 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

Exploratory Time Series Analysis

- Start with a time series plot, i.e., to plot y_t versus t
  Lake Huron Time Series
- 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

Features of Time Series Data

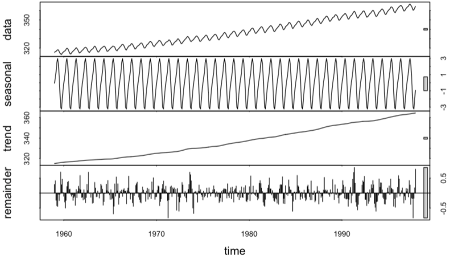
- Trends
  - One can think of trend, μ_t, as continuous changes, usually in the mean, over longer time scales ⇒ *“the essential idea of trend is that it shall be smooth”* - [Kendall, 1973]
  - Typically, the form of the trend is unknown and needs to be estimated. Upon removing the trend, we obtain a 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

Decomposing Time Series into Trend, Seasonality, and Noise

There are two commonly used approaches

- Additive model:

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



- Multiplicative model:

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

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

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

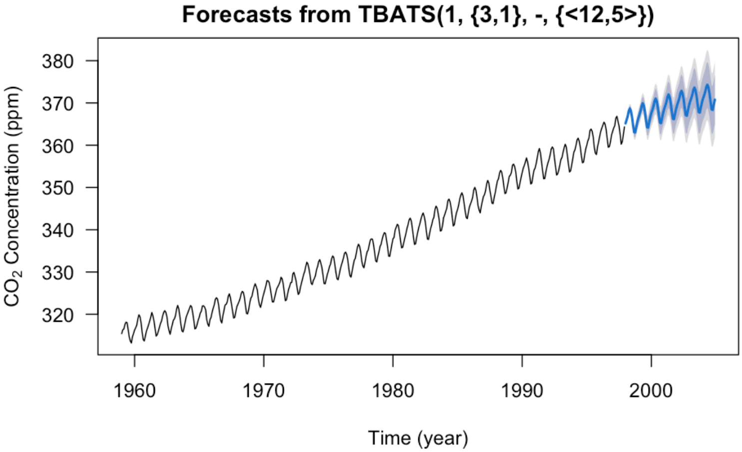
Some Objectives of Time Series Analysis

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 differ years and with a decreasing long-term trend
- The fitted model can be used for further statistical inference, for instant, 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.



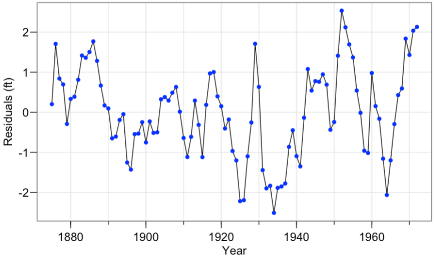
Some Objectives of Time Series Analysis, Cont'd

- **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)

Time Series Models

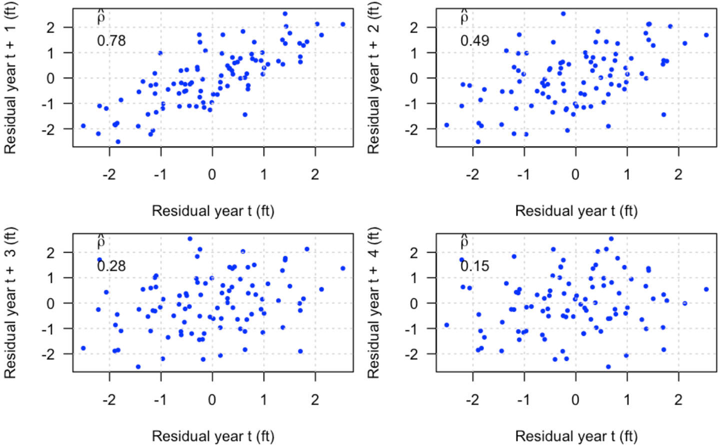
Lake Huron Time Series

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



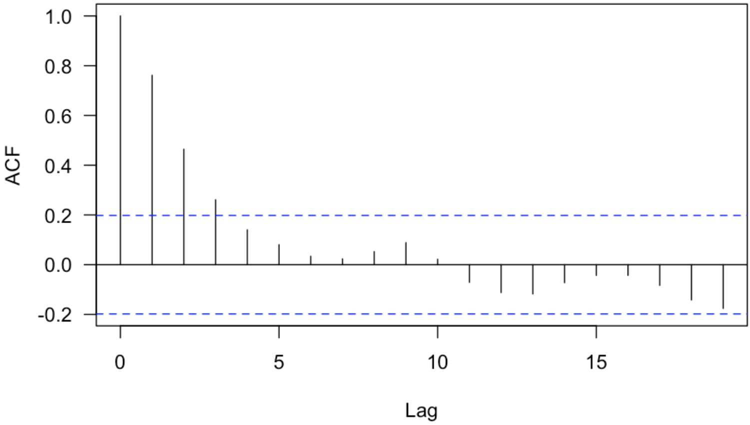
Exploring the Temporal Dependence Structure of $\{\eta_t\}$

$\{\eta_t\}$ exhibit some temporal dependence structure, that is, the nearby (in time) values tend to be more alike than those far part 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

Time Series Models

- 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"

  We will focus on stationarity in means and autocovariances

- 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 $\{\eta_t\}$.

Mean and Autocovariance

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

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

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

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

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

Autocorrelation Function

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

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

It measures the strength of linear association between η_t and $\eta_{t'}$

Properties:

1. $-1 \le \rho(t,t') \le 1, \quad t,t' \in T$
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

Stationary Processes

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":

- $\mathrm{E}[\eta_t] = 0, \quad \forall t \in T$
- $\mathrm{Cov}(\eta_t, \eta_{t'}) = \gamma(t' - t) = \mathrm{Cov}(\eta_{t+s}, \eta_{t'+s})$

$\Rightarrow$ autocorrelation function (ACF):

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

Autoregressive Moving Average (ARMA) Models

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

- 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 Moving Average (ARMA) Models

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

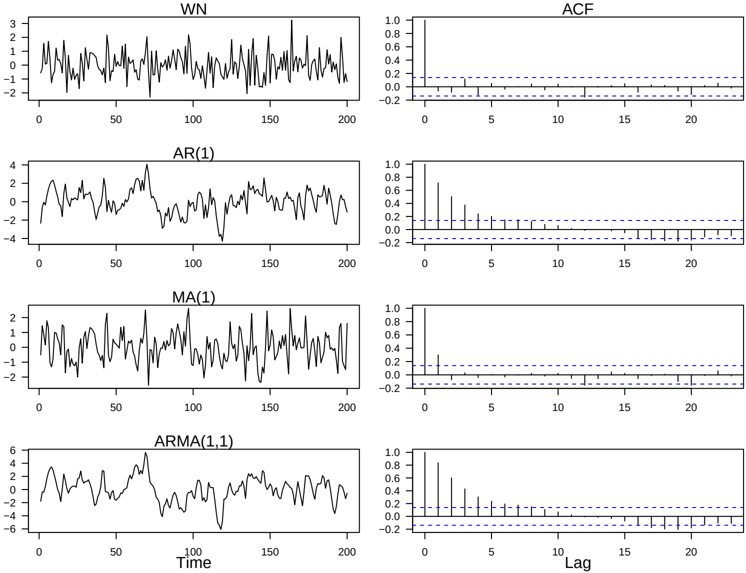
- Moving Average Processes (MA(q)):
  $\eta_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} \cdots + \theta_q Z_{t-q}$
- Autoregressive Processes (AR(p)):
  $\eta_t = \phi_1 \eta_{t-1} + \phi_2 \eta_{t-2} + \cdots + \phi_p \eta_{t-p} + Z_t$

Autoregressive Moving Average (ARMA) Models

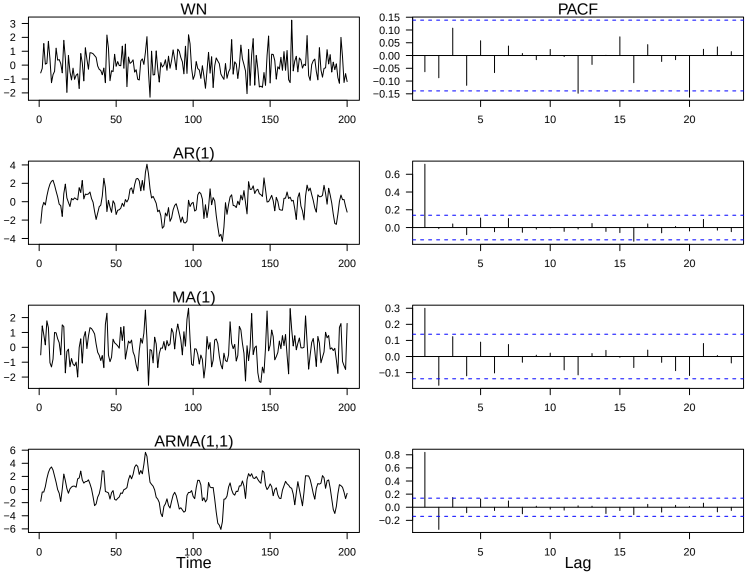
Let $\{Z_t\}$ be independent and identical random variables that follow $\mathrm{N}(0, \sigma^2)$

- 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



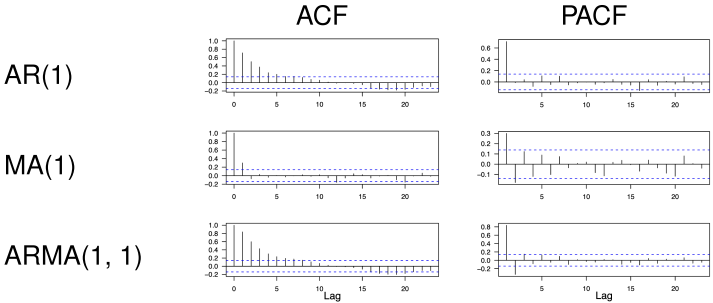
PACF Plots


WN
PACF
AR(1)
MA(1)
ARMA(1,1)
Time
Lag
0
50
100
150
200
5
10
15
20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
0.6
0.4
0.2
0.0
0.3
0.2
0.1
0.0
−0.1
0.8
0.6
0.4
0.2
0.0
−0.2

Identification of ARMA Models using ACF/PACF Plots

Use the ACF and PACF together to identify candidate models. The following table gives some rough guidelines.

	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



Unfortunately, it's not a well-defined process and some guesswork is usually needed

Model Diagnostics: Ljung-Box Test [Ljung and Box, 1978]

We wish to test:

$H_0 : \{e_1, e_2, \cdots, e_T\}$ is an i.i.d. noise sequence $\Rightarrow$ model adequate
$H_1 : H_0$ is false $\Rightarrow$ model not good,

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

Test statistic:

$$Q_{LB} = T(T-2) \sum_{h=1}^{\text{lag}} \frac{\hat{\rho}^2_{\hat{e}}(h)}{T-h} \overset{H_0}{\approx} \chi^2_k,$$

where T is the sample size, $\hat{\rho}_{\hat{e}}(h)$ is the sample ACF at lag h, applied to the residuals of a fitted ARIMA model. The degrees of freedom $k = \text{Lag} - p - q$.

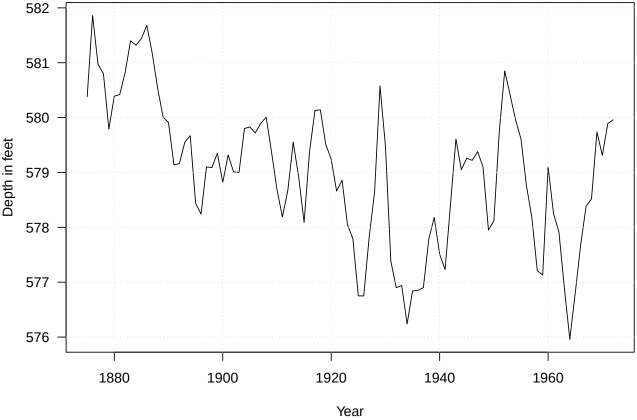
Ljung-Box test can be carried out in `R` using the function `Box.test`

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

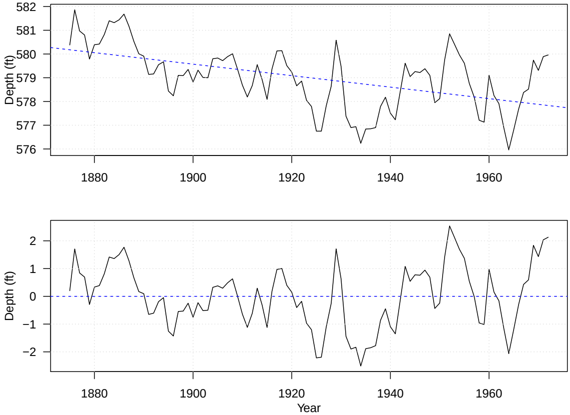


There seems to be a decreasing trend ⇒ need to estimate the trend to get the detrended series

Plots of the Trend and Residuals

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

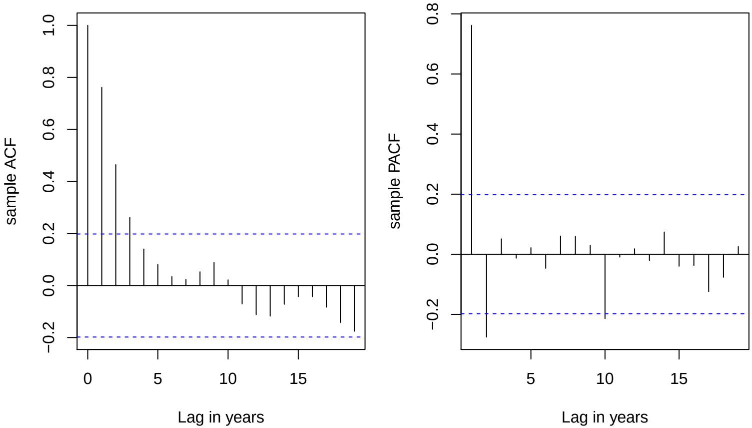
where we **assume** $\mu_t = \alpha + \beta t$, i.e., a linear trend in time



Return

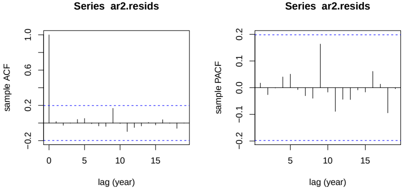
ACF and PACF Plots

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



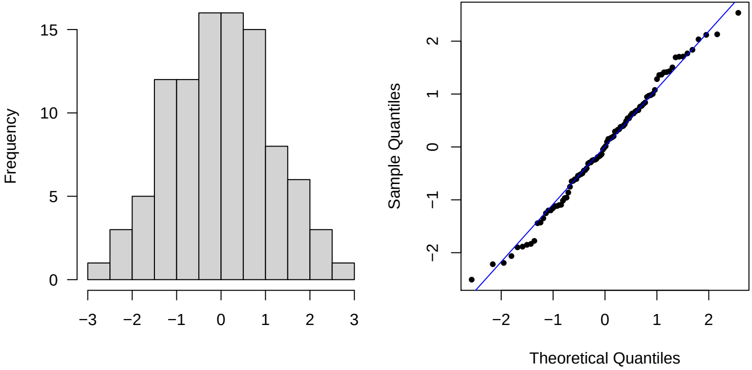
Fitting an AR(2) to the Detrended Time Series

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

Call:
arima(x = deTrend, order = c(2, 0, 0), method = "ML")

Coefficients:
         ar1      ar2  intercept
      1.0047  -0.2919     0.0197
s.e.  0.0977   0.1004     0.2350

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




```
> Box.test(ar2.resids, lag = 5, fitdf = 2, type = "Ljung-Box")

        Box-Ljung test

data:  ar2.resids
X-squared = 0.55962, df = 3, p-value = 0.9056
```

Assessing Normality Assumption for η_t



- Histogram: To compare the shape of the distribution of residuals with the bell-shaped normal density curve
- Q-Q plot: To compare the quantiles of the residual distribution to the quantiles of a normal distribution

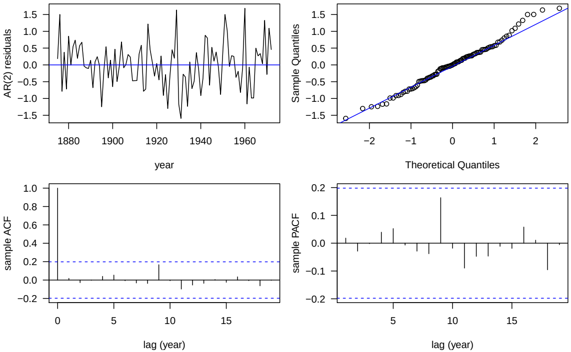
Model Selection via `AIC`

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

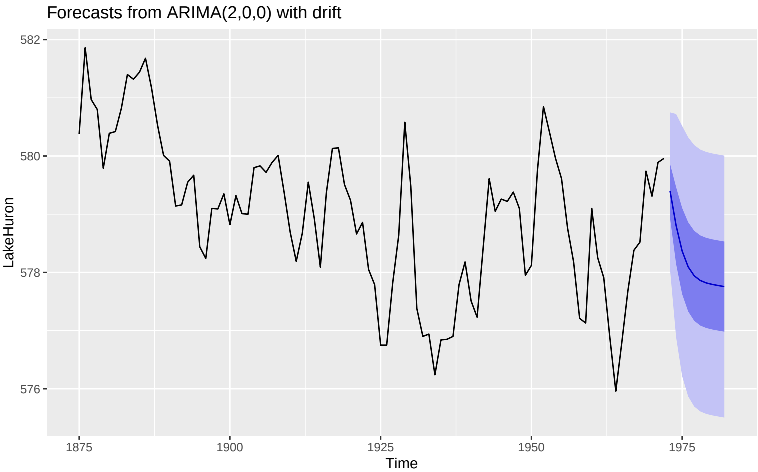
```
> auto.arima(deTrend, trace = T)

 ARIMA(2,0,2) with non-zero mean : 215.0455
 ARIMA(0,0,0) with non-zero mean : 304.222
 ARIMA(1,0,0) with non-zero mean : 216.8388
 ARIMA(0,0,1) with non-zero mean : 235.4585
 ARIMA(0,0,0) with zero mean     : 302.1373
 ARIMA(1,0,2) with non-zero mean : 212.7747
 ARIMA(0,0,2) with non-zero mean : 218.2478
 ARIMA(1,0,1) with non-zero mean : 210.9477
 ARIMA(2,0,1) with non-zero mean : 212.8306
 ARIMA(2,0,0) with non-zero mean : 210.9333
 ARIMA(3,0,0) with non-zero mean : 212.7787
 ARIMA(3,0,1) with non-zero mean : Inf
 ARIMA(2,0,0) with zero mean     : 208.7655
 ARIMA(1,0,0) with zero mean     : 214.7735
 ARIMA(3,0,0) with zero mean     : 210.569
 ARIMA(2,0,1) with zero mean     : 210.6186
 ARIMA(1,0,1) with zero mean     : 208.7891
 ARIMA(3,0,1) with zero mean     : Inf

 Best model: ARIMA(2,0,0) with zero mean
```

Fitting Linear Trend and ARMA in One Step

```
> (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

sigma^2 = 0.476:  log likelihood = -101.2
AIC=212.4   AICc=213.05   BIC=225.32
```



10-Year-Ahead Forecasts



Summary

These slides cover:

- Basic concepts of time series analysis
- A widely used class of models: ARMA
- ARMA model identification, estimation/prediction, inference

`R` functions to know:

- `acf` and `pacf` for identifying candidate models
- `arima` and `Arima` (under the package `forecast`) for model fitting
- `auto.arima` for model selection
- `Box.test` for testing model adequacy
- `forecast` (under the package `forecast`) for generating forecasts and prediction intervals