Non-Linear Models & Time Series Analysis

Cancer Research UK – 19^{th} of July 2017

D.-L. Couturier / M. Dunning / R. Nicholls

What to use and when:

	Multiple regressors	Non-Gaussian error model	Non-linear model	Autocorrellated data
Simple regression				
Multiple regression	✔			
Generalised linear model	✔	✔		
Non-linear model	✔	✔	✔	
Time series analysis				✔

Non-linear models:

Motivating example: *Predicting timber volume of cherry trees*

$$\boldsymbol{y} = \beta_0 \boldsymbol{x}_{\mathbf{1}}^{\beta_1} \boldsymbol{x}_{\mathbf{2}}^{\beta_2} + \boldsymbol{\varepsilon}$$

Response: y = Volume
Predictor: x_1 = Girth
Predictor: x_2 = Height

Can't solve using standard regression approaches.

Instead, use a library that can estimate parameters for non-linear models, e.g. "nls" in R.

Con's:

- May require initial parameter estimates
- May not find globally optimal solution – depends on initial parameter estimates
- May not converge at all
- Slower – iterative approach
- Becomes slower and less reliable as the function becomes more complex

Pro's:

- Allows dealing with a wider class of model functional forms

Non-linear models:

Motivating example: *Predicting timber volume of cherry trees*

$$y = \beta_0 x_1^{\beta_1} x_2^{\beta_2} + \varepsilon$$

Response: y = Volume
Predictor: x_1 = Girth
Predictor: x_2 = Height

Can't solve using standard regression approaches.

Instead, use a library that can estimate parameters for non-linear models, e.g. "nls" in R.

```
Parameters:
        Estimate  Std. Error t value Pr(>|t|)
beta0 0.001449   0.001367   1.060 0.298264
beta1 1.996921   0.082077  24.330  < 2e-16 ***
beta2 1.087647   0.242159   4.491 0.000111 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.533 on 28 degrees of freedom

Number of iterations to convergence: 5
Achieved convergence tolerance: 8.255e-07
```

AIC = 150.4

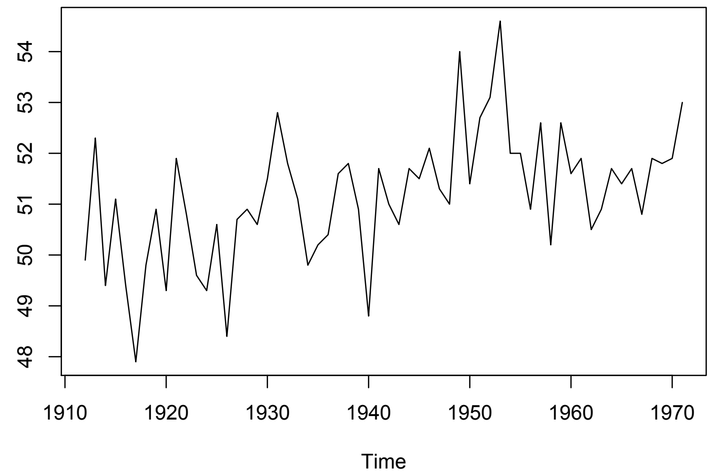
```
Parameters:
        Estimate  Std. Error t value Pr(>|t|)
beta1  2.27405   0.12967   17.54  < 2e-16 ***
beta2 -0.58432   0.08242   -7.09 8.44e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.216 on 29 degrees of freedom

Number of iterations to convergence: 10
Achieved convergence tolerance: 8.673e-06
```

AIC = 181.1

Note: poor parameter interpretation

Time series analysis

A time series is a process in which a given observation depends on other datapoints in the same series.



Linear regression models:

- Response variable (y)
- Independent variables (x)

Time series:

- Single process (y)

Idea:

- Exploit correlation in order to understand and model data
- Potentially forecast likelihood of future events

Time series analysis

When analysing time series, we are interested in how two values in the series – separated by k time-steps – affect each other.

k^th autocovariance:

$$\gamma_k = E(y_t - \mu)(y_{t-k} - \mu)$$

Average covariance between pairs of values that are k time steps apart in the series.

Since these are dependent on the scale of the process, these need to be standardised:

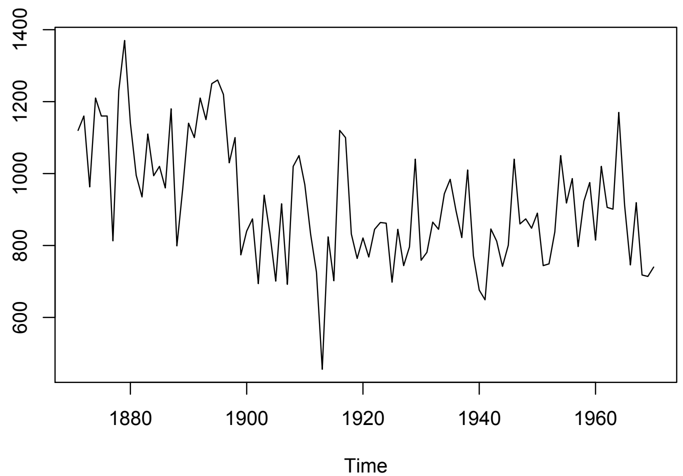
k^th autocorrelation: $$\rho_k = \frac{\gamma_k}{\gamma_0}$$

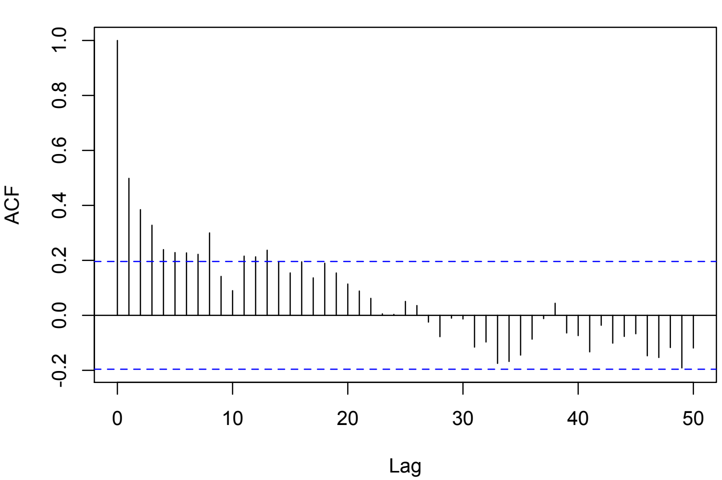
The autocorrelation function is useful for characterising time series.

Time series analysis

Autocorrelation function:

Nile annual flow:

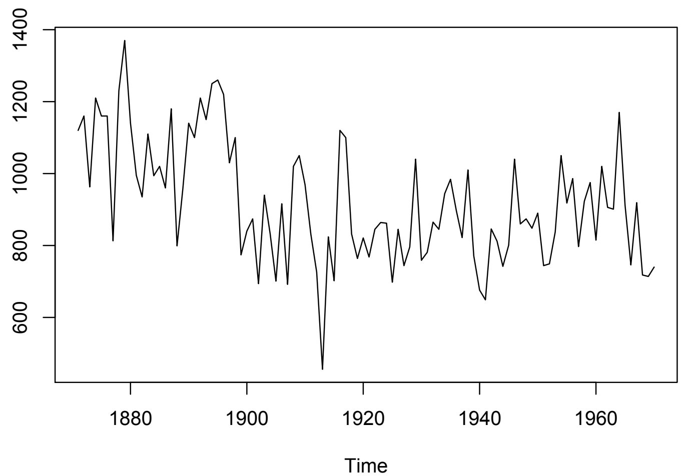


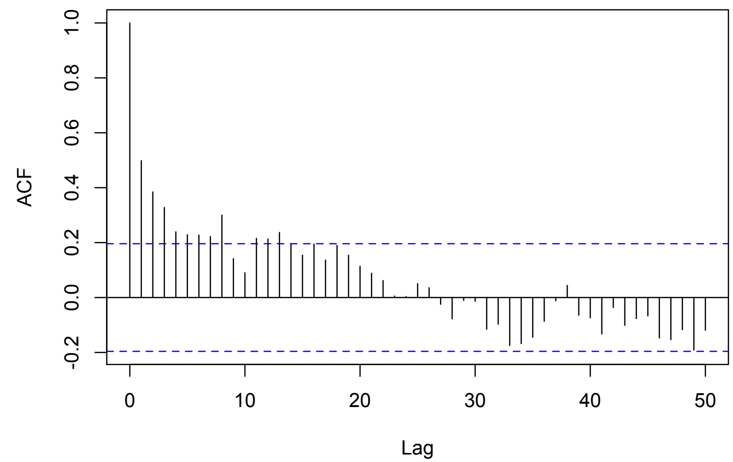


Time series analysis

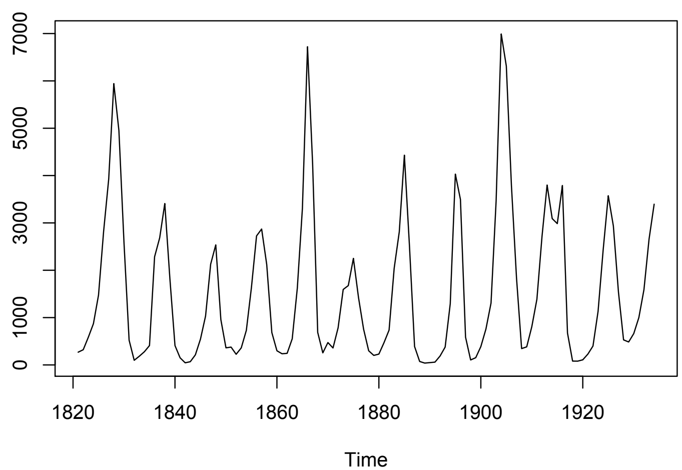
Autocorrelation function:

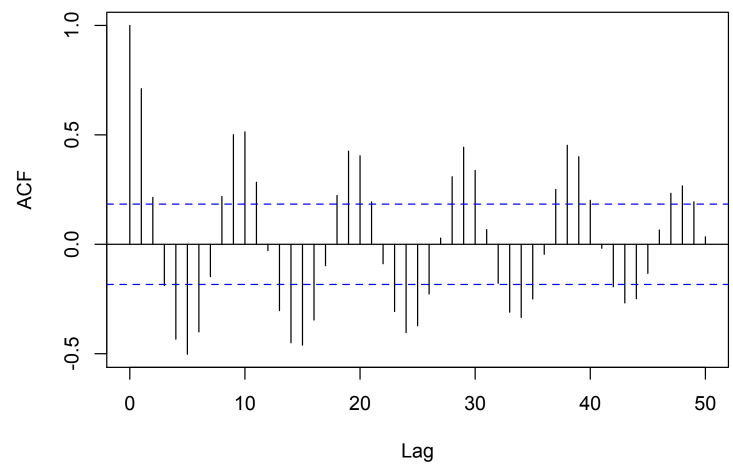
Nile annual flow:





Lynx trappings:





Time series analysis

Autoregressive (AR) time series models:

AR(1): $$y_t = c + \varphi_1 y_{t-1} + \varepsilon_t$$

Time series analysis

Autoregressive (AR) time series models:

AR(1): $y_t = c + \varphi_1 y_{t-1} + \varepsilon_t$

AR(2): $y_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \varepsilon_t$

Time series analysis

Autoregressive (AR) time series models:

AR(1): $$y_t = c + \varphi_1 y_{t-1} + \varepsilon_t$$

AR(2): $$y_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \varepsilon_t$$

AR(p): $$y_t = c + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \varepsilon_t$$

Similarities to multiple regression model, except for the dependencies
Parameters estimated using least squares or maximum likelihood

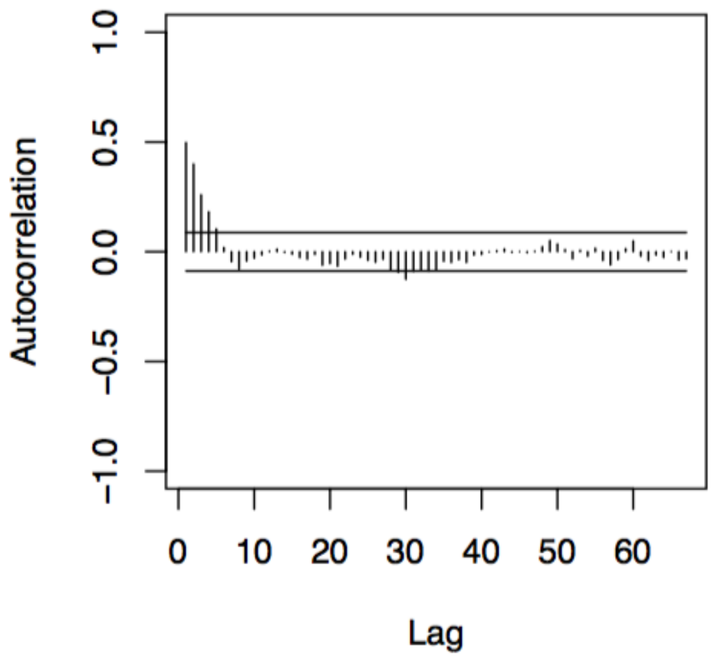
Assumptions:

- Independent Gaussian errors
- Covariance stationary process (trend doesn't change over time)

Time series analysis

Autoregressive (AR) time series models:

AR(2) with c=0, φ_1=0.4 and φ_2=0.2



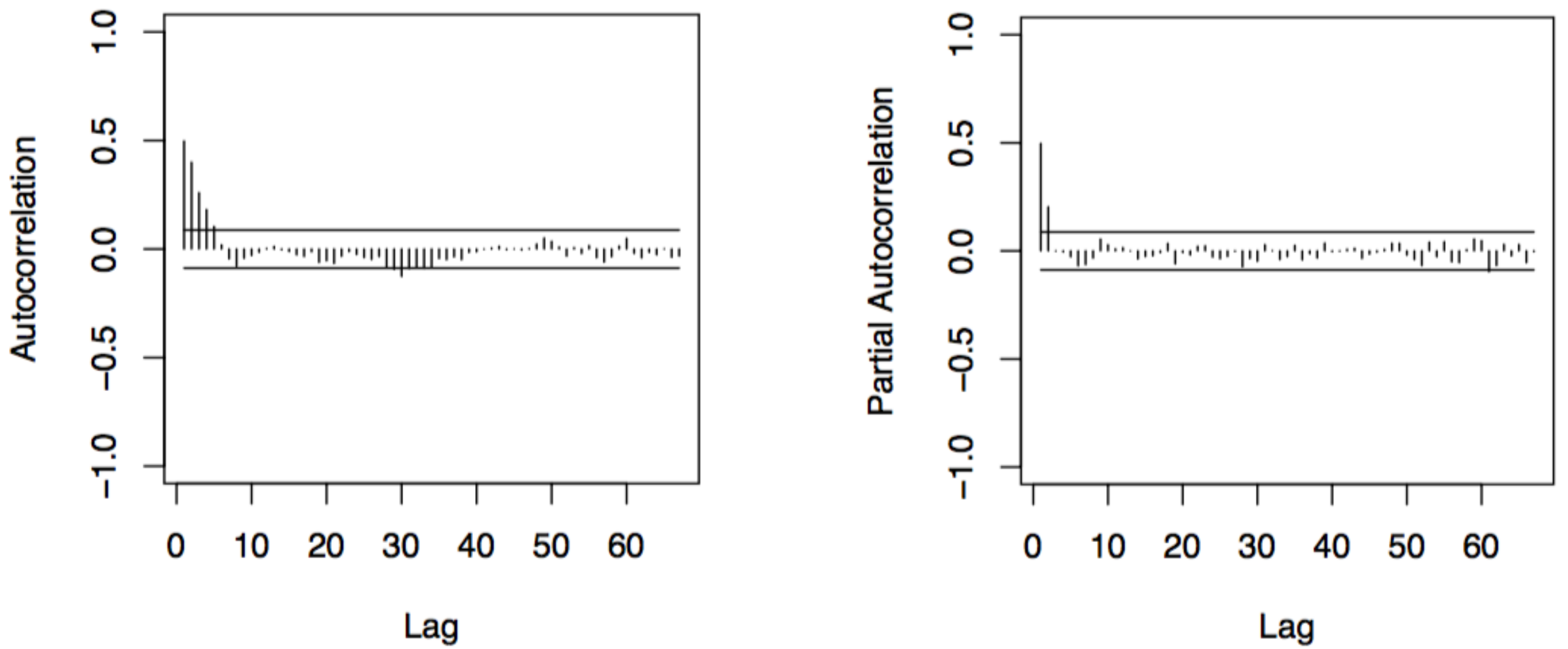
How to interpret ACF?

- Positive parameters: ACF should decay, not oscillate.
- Should decay gradually until within the confidence interval, then stay there.
- Can't infer order…

Time series analysis

Autoregressive (AR) time series models:

AR(2) with $c=0$, $\varphi_1=0.4$ and $\varphi_2=0.2$



Partial autocorrelation function: $\alpha(p) = \varphi_p$ from a AR(p) model

Parsimonious modelling:

- First try AR(1), then AR(2), etc. until $H_0 : \alpha(p) = 0$ is not rejected.
- Failure to reject leads us to conclude that AR(p-1) is more appropriate than AR(p).

Time series analysis

Moving Average (MA) time series models:

MA(1): $$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

MA(2): $$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}$$

MA(q): $$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

Unlike multiple regression model there are multiple error terms
However, the current state is only ever dependent on known no. of previous states

Since the current state only depends on the previous q states,
the ACF should suddenly drop to zero, unlike AR(p) processes

Time series analysis

More general models:

ARMA(p,q): $$y_t = c + \underbrace{\varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p}}_{\text{AR(p)}} + \varepsilon_t + \underbrace{\theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}}_{\text{MA(q)}}$$

Time series analysis

More general models:

ARMA(p,q): $$y_t = c + \underbrace{\varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p}}_{\text{AR(p)}} + \varepsilon_t + \underbrace{\theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}}_{\text{MA(q)}}$$

ARIMA(p,1,q): $x_t = y_t - y_{t-1}$ then model as ARMA(p,q)

Time series analysis

More general models:

ARMA(p,q): $$y_t = c + \underbrace{\varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p}}_{\text{AR(p)}} + \varepsilon_t + \underbrace{\theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}}_{\text{MA(q)}}$$

ARIMA(p,1,q): $x_t = y_t - y_{t-1}$ then model as ARMA(p,q)

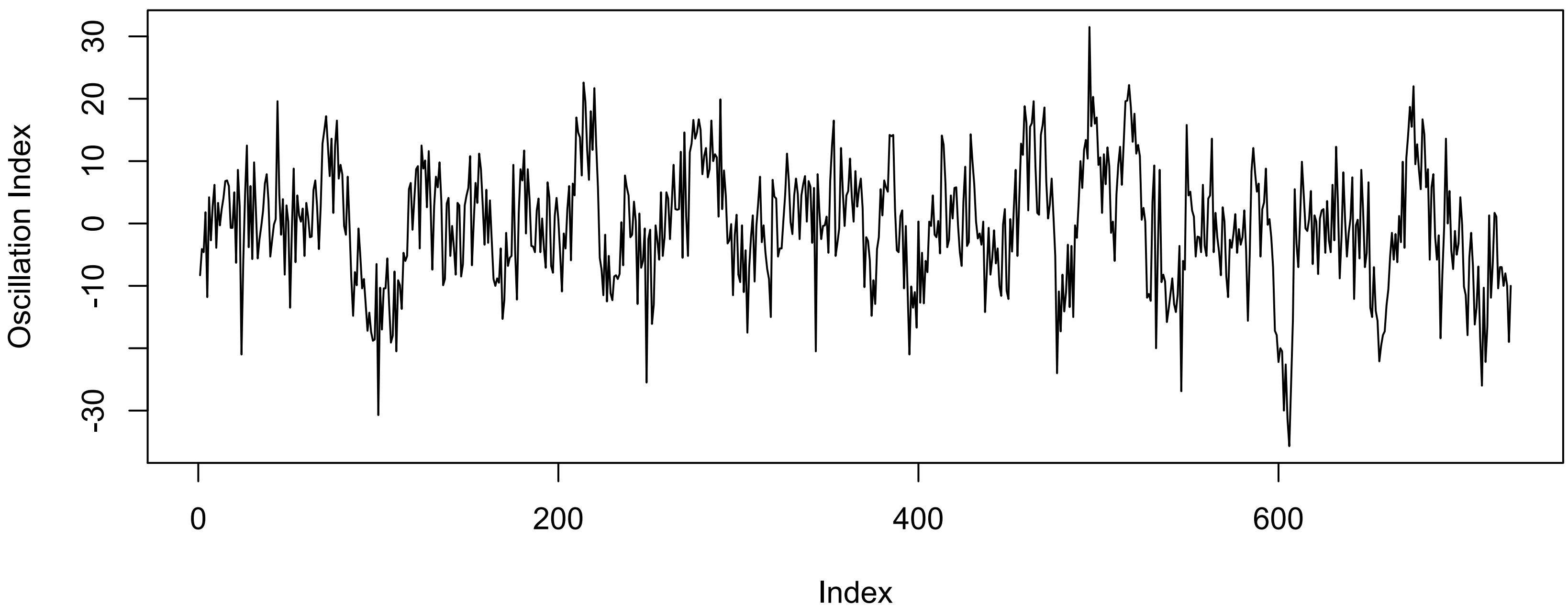
ARIMA(p,d,q): $x_t = \nabla^d y_t$ take d^th order differences

Considering ARIMA models can be a useful "transformation" if assumptions are violated

Time series analysis

Example: Monthly Southern Oscillation Index

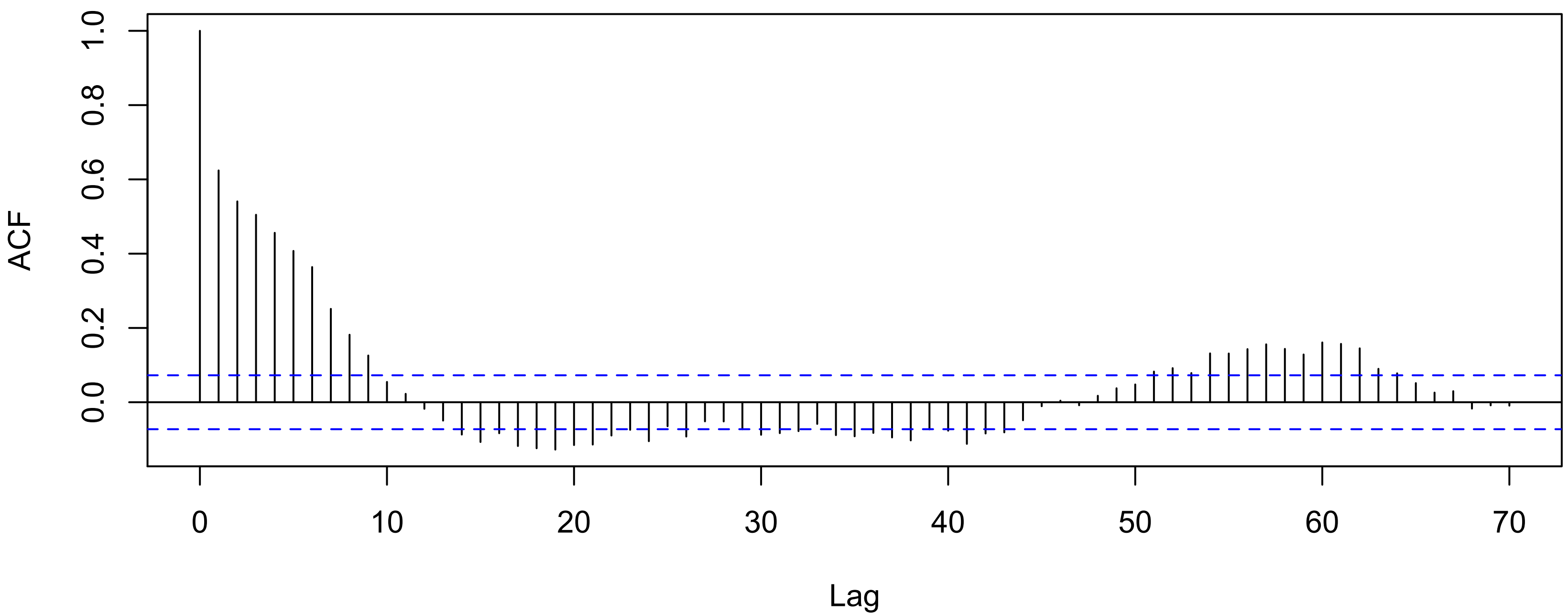
Monthly difference in sea-surface air pressure between Darwin and Tahiti



Time series analysis

Example: Monthly Southern Oscillation Index

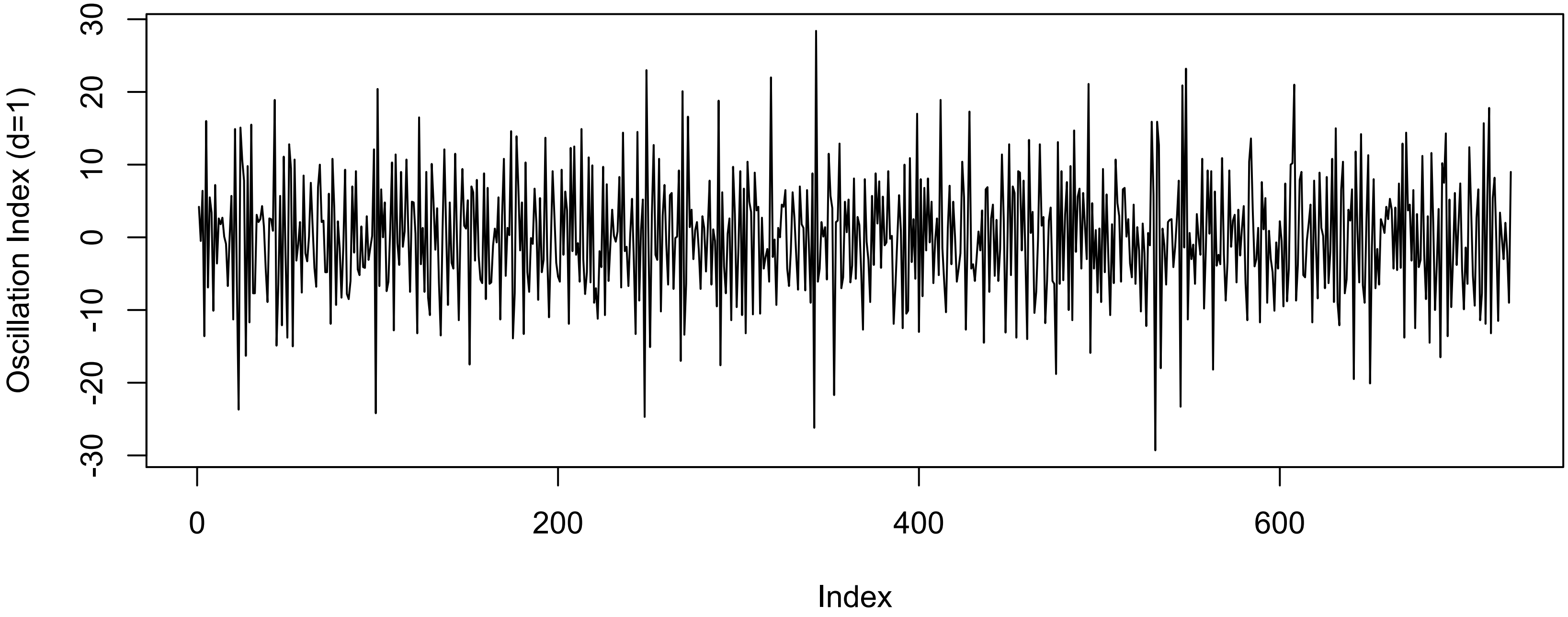
Monthly difference in sea-surface air pressure between Darwin and Tahiti



Time series analysis

Example: Monthly Southern Oscillation Index

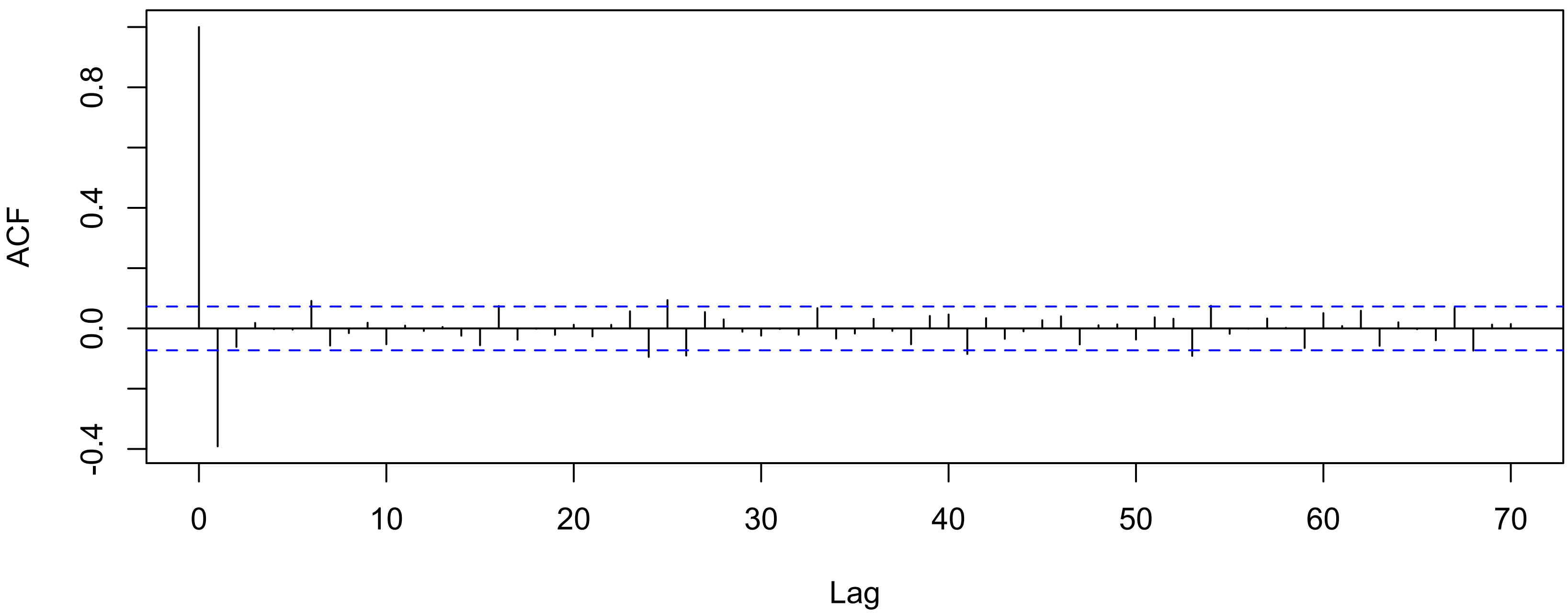
Monthly difference in sea-surface air pressure between Darwin and Tahiti



Time series analysis

Example: Monthly Southern Oscillation Index

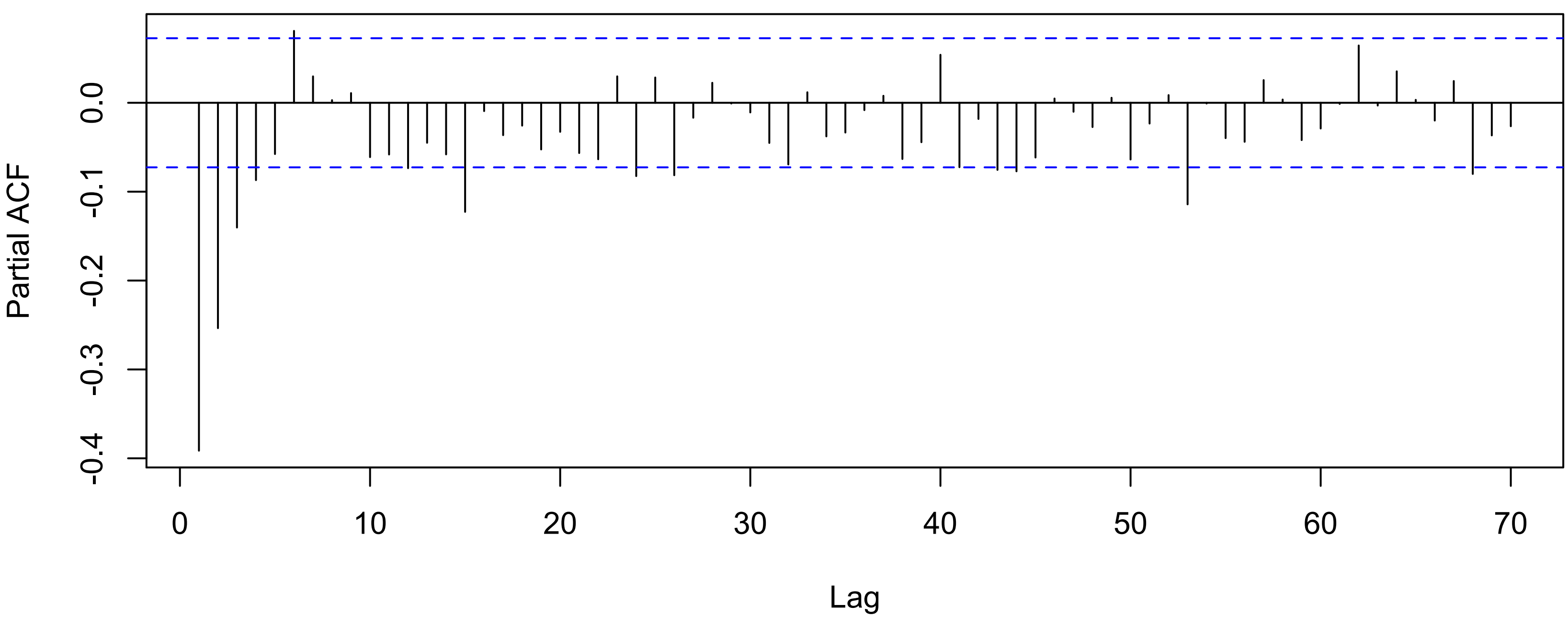
Monthly difference in sea-surface air pressure between Darwin and Tahiti



Time series analysis

Example: Monthly Southern Oscillation Index

Monthly difference in sea-surface air pressure between Darwin and Tahiti



Time series analysis

Example: Monthly Southern Oscillation Index

Monthly difference in sea-surface air pressure between Darwin and Tahiti

Try ARIMA(0,1,1) model:

```
arima(x = x$Index, order = c(0, 1, 1))

Coefficients:
          ma1
      -0.5579
s.e.   0.0308

sigma^2 estimated as 52.94:  log likelihood = -2477.98,  aic = 4959.96
```