Lecture 7

AR Models

Colin Rundel
02/08/2017

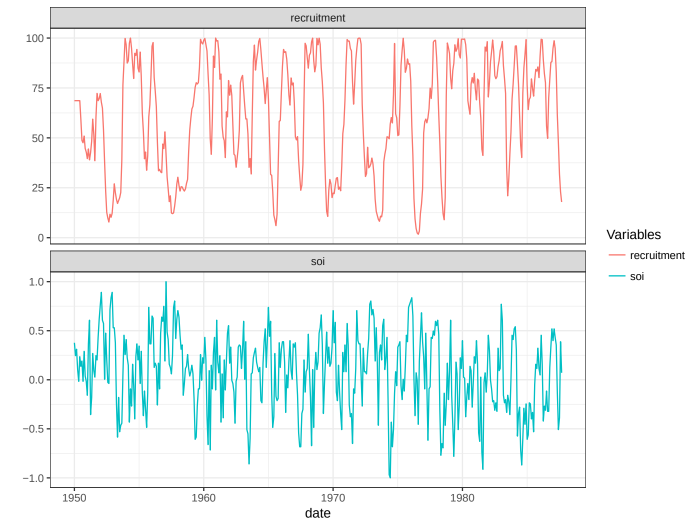
Lagged Predictors and CCFs

Southern Oscillation Index & Recruitment

The Southern Oscillation Index (SOI) is an indicator of the development and intensity of El Niño (negative SOI) or La Niña (positive SOI) events in the Pacific Ocean. These data also included the estimate of “recruitment”, which indicate fish population sizes in the southern hemisphere.

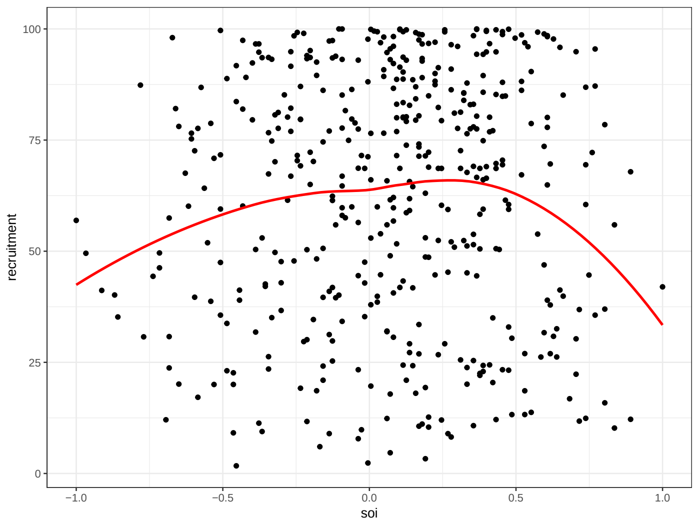
```
## # A tibble: 453 × 3
##        date    soi recruitment
##       <dbl>  <dbl>       <dbl>
## 1  1950.000  0.377       68.63
## 2  1950.083  0.246       68.63
## 3  1950.167  0.311       68.63
## 4  1950.250  0.104       68.63
## 5  1950.333 -0.016       68.63
## 6  1950.417  0.235       68.63
## 7  1950.500  0.137       59.16
## 8  1950.583  0.191       48.70
## 9  1950.667 -0.016       47.54
## 10 1950.750  0.290       50.91
## # ... with 443 more rows
```

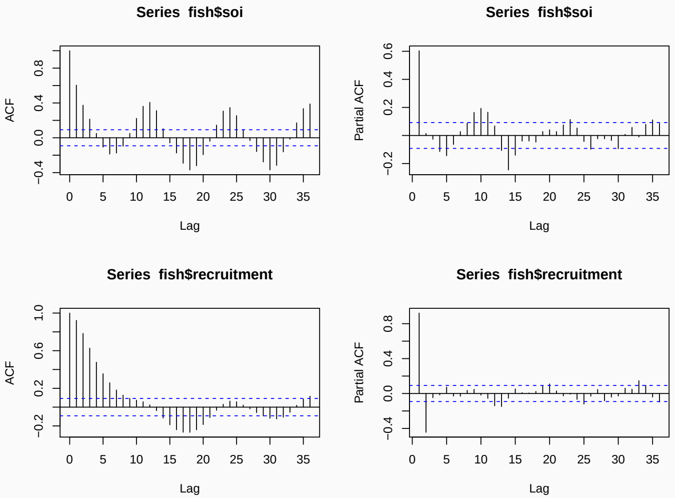
Time series



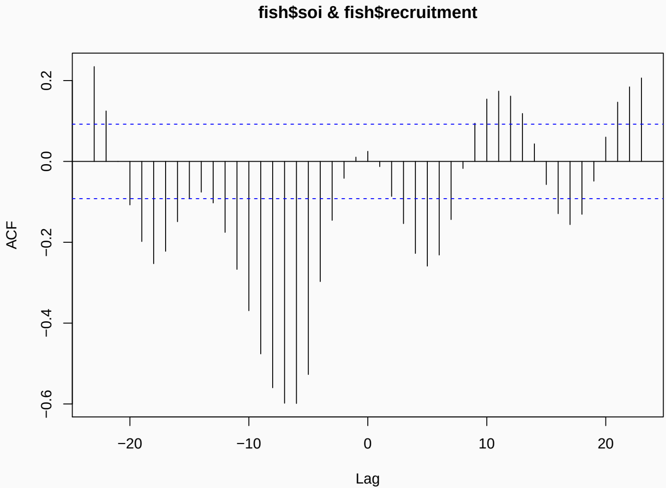
recruitment
100
75
50
25
0
soi
1.0
0.5
0.0
−0.5
−1.0
1950
1960
1970
1980
date
Variables
recruitment
soi

Relationship?

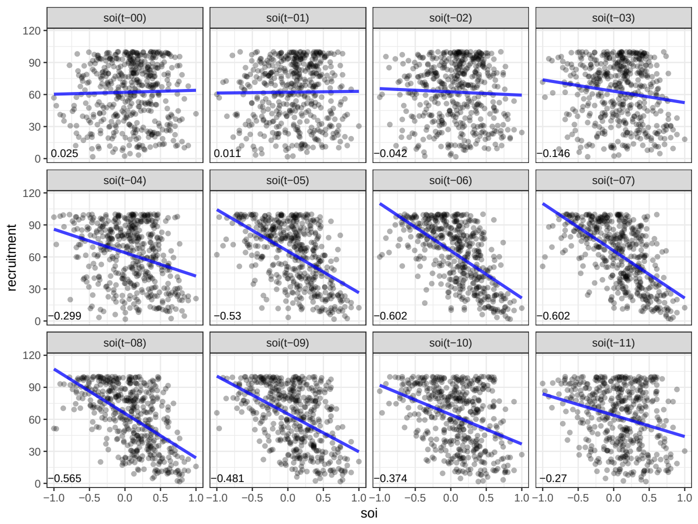



Series fish$soi
ACF
Lag
Series fish$soi
Partial ACF
Lag
Series fish$recruitment
ACF
Lag
Series fish$recruitment
Partial ACF
Lag

Cross correlation function



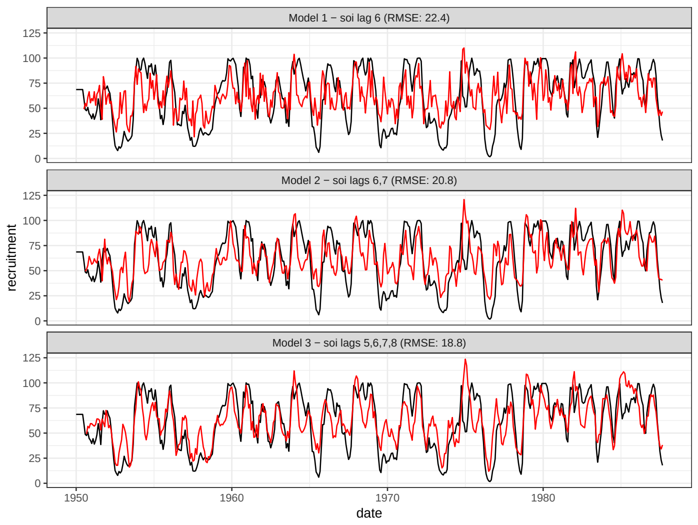
Cross correlation function - Scatter plots



Model

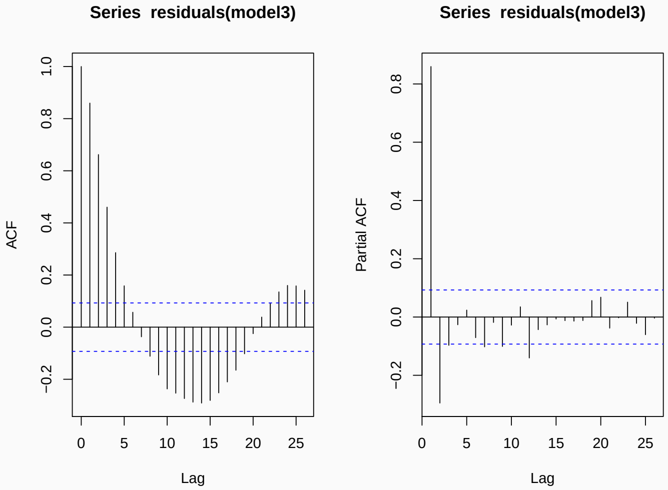
```
##
## Call:
## lm(formula = recruitment ~ lag(soi, 5) + lag(soi, 6) + lag(soi,
##     7) + lag(soi, 8), data = fish)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -72.409 -13.527   0.191  12.851  46.040
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  67.9438     0.9306  73.007  < 2e-16 ***
## lag(soi, 5) -19.1502     2.9508  -6.490 2.32e-10 ***
## lag(soi, 6) -15.6894     3.4334  -4.570 6.36e-06 ***
## lag(soi, 7) -13.4041     3.4332  -3.904 0.000109 ***
## lag(soi, 8) -23.1480     2.9530  -7.839 3.46e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.93 on 440 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.5539, Adjusted R-squared:  0.5498
## F-statistic: 136.6 on 4 and 440 DF,  p-value: < 2.2e-16
```

Prediction



Model 1 − soi lag 6 (RMSE: 22.4)
Model 2 − soi lags 6,7 (RMSE: 20.8)
Model 3 − soi lags 5,6,7,8 (RMSE: 18.8)
recruitment
125
100
75
50
25
0
1950
1960
1970
1980
date

Residual ACF - Model 3

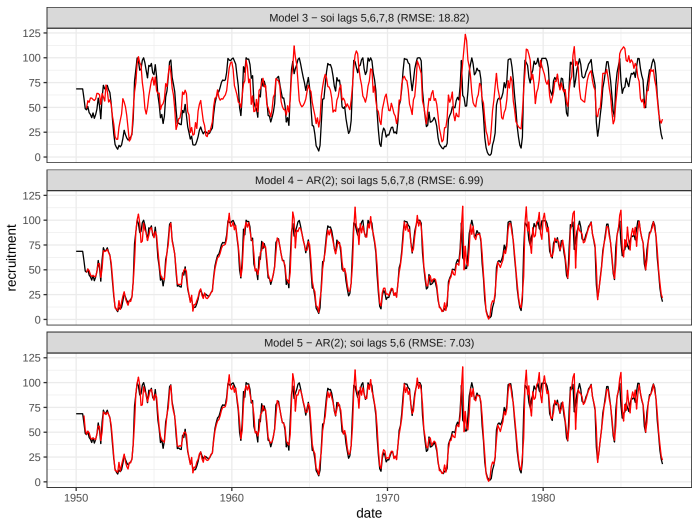


```
##
## Call:
## lm(formula = recruitment ~ lag(recruitment, 1) + lag(recruitment,
##     2) + lag(soi, 5) + lag(soi, 6) + lag(soi, 7) + lag(soi, 8),
##     data = fish)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.996  -2.892   0.103   3.117  28.579
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          10.25007    1.17081   8.755  < 2e-16 ***
## lag(recruitment, 1)   1.25301    0.04312  29.061  < 2e-16 ***
## lag(recruitment, 2)  -0.39961    0.03998  -9.995  < 2e-16 ***
## lag(soi, 5)         -20.76309    1.09906 -18.892  < 2e-16 ***
## lag(soi, 6)           9.71918    1.56265   6.220 1.16e-09 ***
## lag(soi, 7)          -1.01131    1.31912  -0.767   0.4437
## lag(soi, 8)          -2.29814    1.20730  -1.904   0.0576 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.042 on 438 degrees of freedom
##   (8 observations deleted due to missingness)
```

Autoregessive model 2

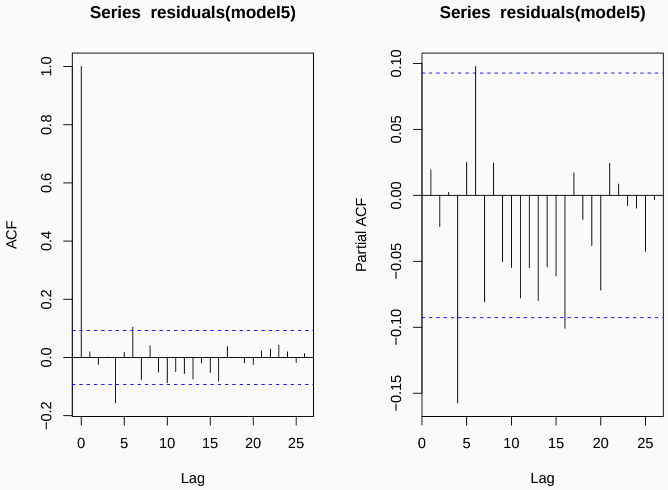
```
##
## Call:
## lm(formula = recruitment ~ lag(recruitment, 1) + lag(recruitment,
##     2) + lag(soi, 5) + lag(soi, 6), data = fish)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.786  -2.999  -0.035   3.031  27.669
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           8.78498    1.00171   8.770  < 2e-16 ***
## lag(recruitment, 1)   1.24575    0.04314  28.879  < 2e-16 ***
## lag(recruitment, 2)  -0.37193    0.03846  -9.670  < 2e-16 ***
## lag(soi, 5)         -20.83776    1.10208 -18.908  < 2e-16 ***
## lag(soi, 6)           8.55600    1.43146   5.977 4.68e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.069 on 442 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.9375, Adjusted R-squared:  0.937
## F-statistic:  1658 on 4 and 442 DF,  p-value: < 2.2e-16
```

Prediction



Model 3 − soi lags 5,6,7,8 (RMSE: 18.82)
Model 4 − AR(2); soi lags 5,6,7,8 (RMSE: 6.99)
Model 5 − AR(2); soi lags 5,6 (RMSE: 7.03)
recruitment
0
25
50
75
100
125
1950
1960
1970
1980
date

Residual ACF - Model 5



Non-stationarity

Non-stationary models

> *All happy families are alike; each unhappy family is unhappy in its own way. - Tolstoy, Anna Karenina*

This applies to time series models as well, just replace happy family with stationary model.

Non-stationary models

> *All happy families are alike; each unhappy family is unhappy in its own way. - Tolstoy, Anna Karenina*

This applies to time series models as well, just replace happy family with stationary model.

A simple example of a non-stationary time series is a trend stationary model

$$y_t = \mu_t + w_t$$

where μ_t denotes the trend and w_t is a stationary process.

Non-stationary models

> *All happy families are alike; each unhappy family is unhappy in its own way. - Tolstoy, Anna Karenina*

This applies to time series models as well, just replace happy family with stationary model.

A simple example of a non-stationary time series is a trend stationary model

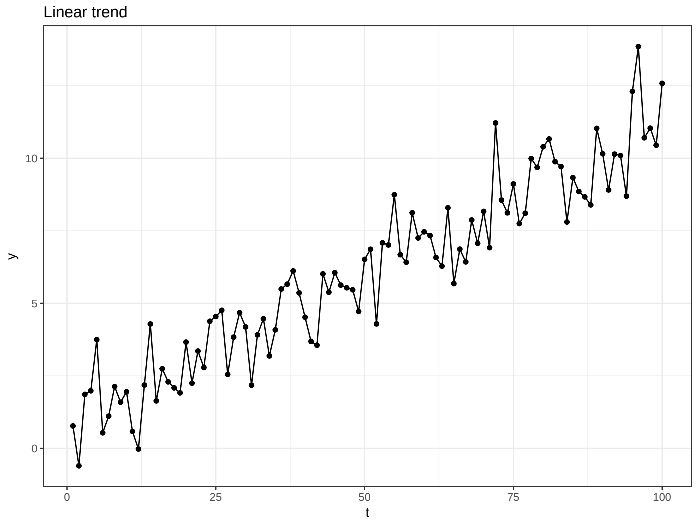
$$y_t = \mu_t + w_t$$

where μ_t denotes the trend and w_t is a stationary process.

We've already been using this approach, since it is the same as estimating μ_t via regression and then examining the residuals ($\hat{w}_t = y_t - \hat{mu}_t$) for stationarity.

Linear trend model

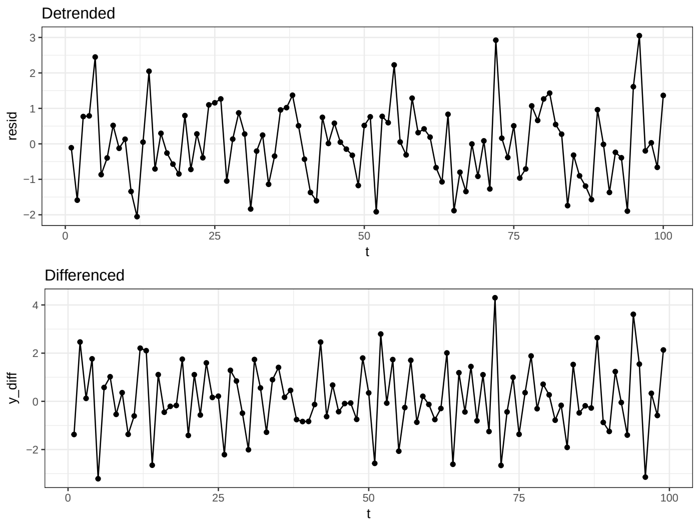
Lets imagine a simple model where $y_t = \delta + \phi t + w_t$ where δ and ϕ are constants and $w_t \sim \mathcal{N}(0, \sigma_w^2)$.



Differencing

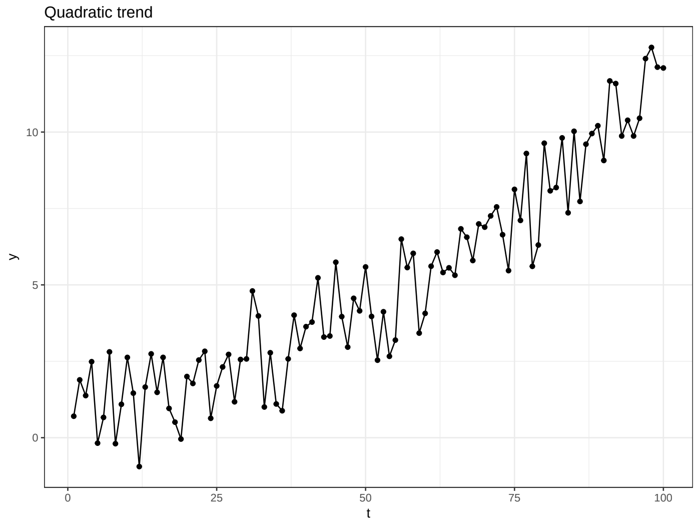
An alternative approach to what we have seen is to examine the differences of your response variable, specifically $y_t - y_{t-1}$.

Detrending vs Difference

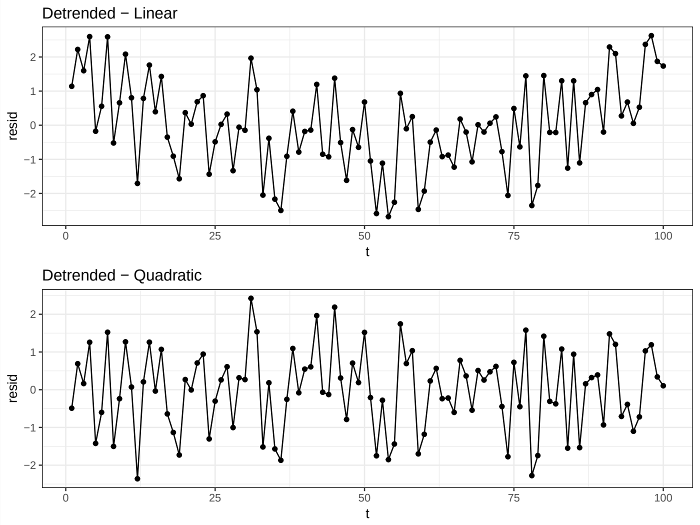


Quadratic trend model

Lets imagine another simple model where $y_t = \delta + \phi t + \gamma t^2 + w_t$ where δ, ϕ, and γ are constants and $w_t \sim \mathcal{N}(0, \sigma_w^2)$.



Detrending

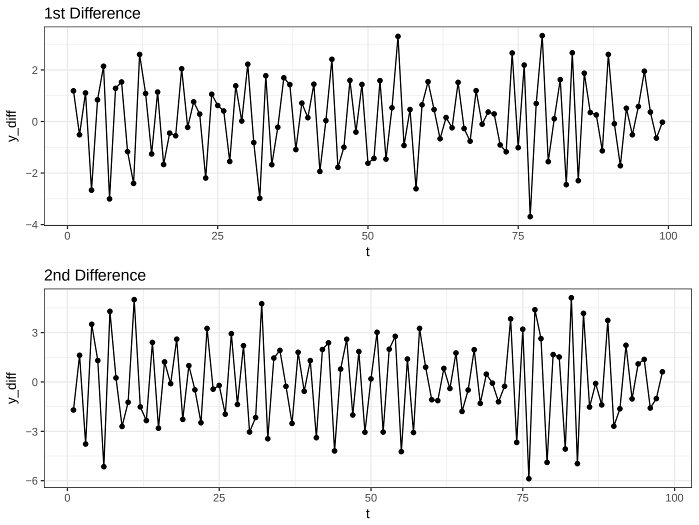


Detrended − Linear
resid
2
1
0
−1
−2
0
25
50
75
100
t
Detrended − Quadratic
resid
2
1
0
−1
−2
0
25
50
75
100
t

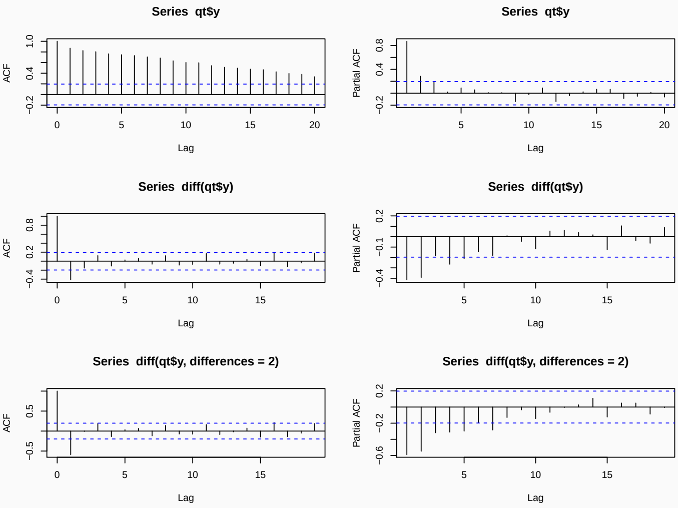
2nd order differencing

Let $d_t = y_t - y_{t-1}$ be a first order difference then $d_t - d_{t-1}$ is a 2nd order difference.

Differencing



Differencing - ACF

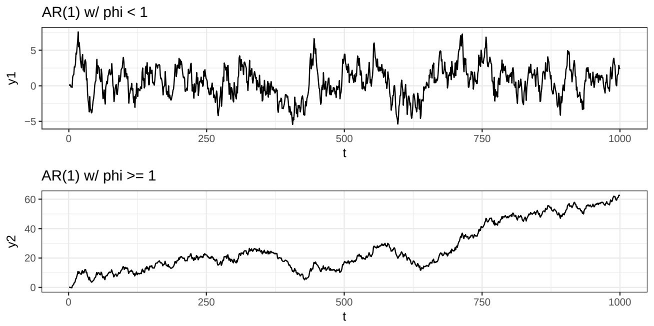


AR Models

AR(1)

Last time we mentioned a random walk with trend process where $y_t = \delta + y_{t-1} + w_t$. The AR(1) process is a slight variation of this where we add a coefficient in front of the y_{t-1} term.

$$AR(1): \quad y_t = \delta + \phi \, y_{t-1} + w_t$$

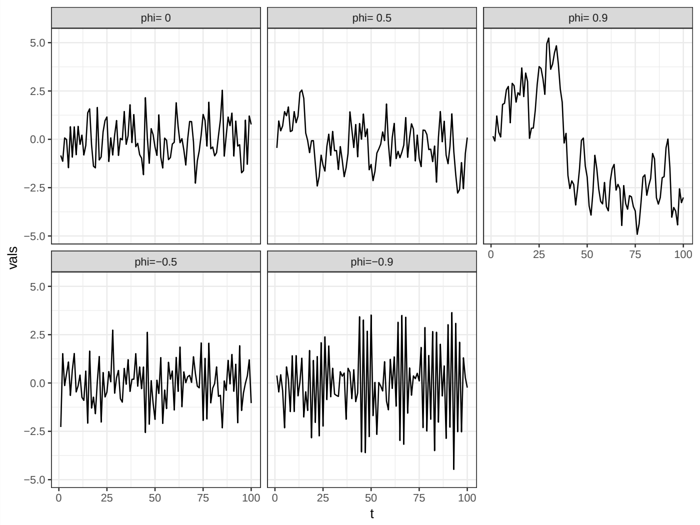


Lets rewrite the AR(1) without any autoregressive terms

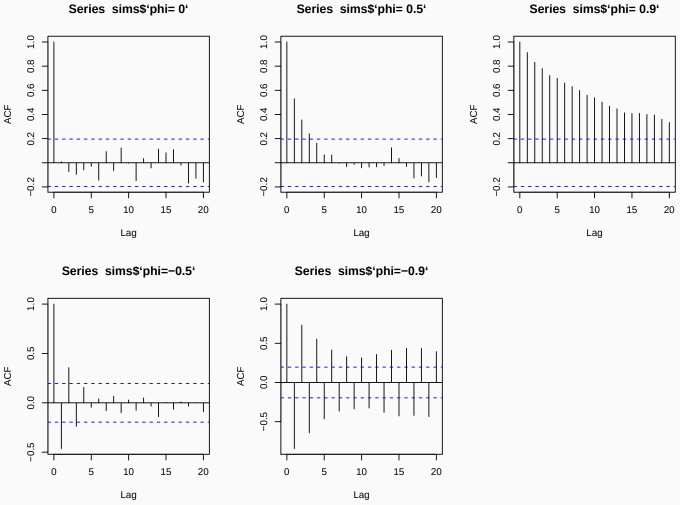
Differencing

Once again we can examine differences of the response variable $y_t - y_{t-1}$ to attempt to achieve stationarity,s

Identifying AR(1) Processes



Identifying AR(1) Processes - ACFs



AR(p) models

We can easily generalize from an AR(1) to an AR(p) model by simply adding additional autoregressive terms to the model.

$$
\begin{aligned}
AR(p): \quad y_t &= \delta + \phi_1 \, y_{t-1} + \phi_2 \, y_{t-2} + \cdots + \phi_p \, y_{t-p} + w_t \\
&= \delta + w_t + \sum_{i=1}^{p} \phi_i \, y_{t-i}
\end{aligned}
$$

More on these next time.