Computational
Social Science

Modeling Temporal Data .II

Roberto Cerina

21.03.2024

University of Amsterdam

Stationarity

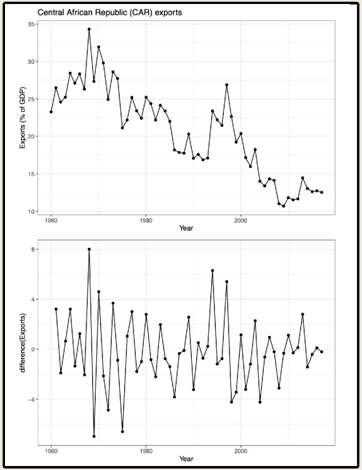
- ✏ ARMA models assume the underlying time-series is **stationary**...
- ✏ Statistical properties do not change over time...
- ✏ ...this is typically not the case !
- ✖ Problems with **non-stationarity**:
  - ✏ non-stationarity implies I cannot generalise out of my 'temporal window' of observations...
  - ✏ I cannot reliably make inference or forecast future values...
  - ✏ say I observe a non-stationary series $\{y_0,...,y_k\}$, and estimate an AR(1) model...
    - fitting an AR(1) model to a subsequent series $\{y_z,...,y_{k+z}\}$ (values from the same series but observed at a later stage) will yield different values of $\hat{\beta}$ and $\hat{\sigma}$;
    - The $\hat{\beta}$ and $\hat{\sigma}$ I can estimate from my sample will be **biased** (different from the true values) and **inconsistent** (increasing sample-size is not guaranteed to get them any closer to the true values).

Differencing

- A time-series which is non-stationary can be **Differenced** to make it more stationary;
- This procedure is not guaranteed to produce a stationary series, but can do so to remove trends or seasonal variation.
- ∇ is called the differencing operator, and it works as follows: $\nabla_i y_t = y_t - y_{t-i}$
- ∇_1 is called 'first-differencing', and is typically used to remove obvious trends from the data.
- The order of differencing is indexed by *i* in the model:

$$ARIMA(p, i, q)$$

Differencing



[0] https://www.stat.berkeley.edu/~ryantibs/timeseries-f23/lectures/arima.pdf

Dickey Fuller Procedure

- We need a mechanism to test whether a series is stationary – visual inspection is too ambiguous...
- we can check whether our time-series is stationary via the **Augmented Dickey Fuller** (ADF) procedure;
- Simple DF develops as follows:
  - We know the random walk model is non-stationary - so if we fit an AR(1) model

    $y_t = \beta_1 y_{t-1} + \epsilon_t$

    and β_1 is statistically indistinguishable from 1, we have a non-stationary time-series;
  - ADF test works with the first-difference of the series - subtract $y_t - 1$ from each side:

    $y_t - y_{t-1} = (\beta_1 - 1)y_{t-1} + \epsilon_t$
    $\nabla_1 y_t = \delta y_{t-1} + \epsilon_t$
  - So in this formulation if δ is not distinguishable from 0 (statistically insignificant), then we have non-stationarity.

Dickey Fuller Procedure: Intuition

"If the series y is stationary ... it has a tendency to return to a constant mean.

Therefore, large values will tend to be followed by smaller values (negative changes), and small values by larger values (positive changes).

Accordingly, the level of the series y_{t-1} will be a significant predictor of next period's change, and will have a negative coefficient."[1]

[1] `https://en.wikipedia.org/wiki/Dickey-Fuller_test`

Dickey Fuller Limitations

- Simple DF does not account for **serial (temporal) correlation in the error terms**;
- Its presence violates the assumption of independent errors required for valid regression analysis…
- This can lead to misleading test statistics and incorrect conclusions about the presence of a unit root.

Augmented Dickey Fuller Procedure

- The Augmented DF (ADF) procedure includes lagged differences of the series as controls.
- This accounts for higher order autoregressive processes, leading to residual autocorrelation.
- The equation is:

$$\nabla_1 y_t = \alpha + \beta t + \gamma y_{t-1} + \nabla_1 y_{t-1} + \cdots + \delta_{p-1} \nabla_1 y_{t-p+1} + \epsilon_t$$

Augmented Dickey Fuller Limitations

- Note: ADF could still (rarely) suggest that processes such as 'oscillating explosive' ($\beta < -1$), and 'explosive' ($\beta > 1$) do not produce unit-roots, and hence allow for non-stationary time series…
- These instances are marginal, degenerate cases and hardly ever arise in practice – be thoughtful in analysing the output of the DF procedure !

Seasonal ARIMA Models

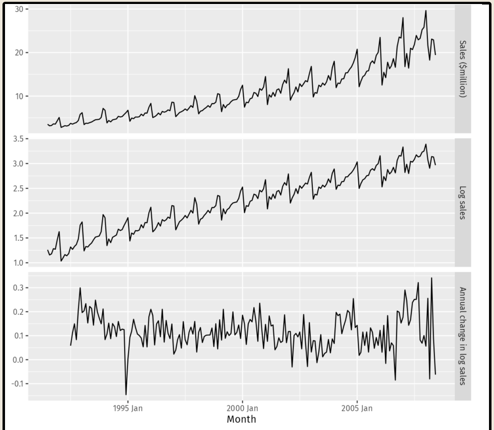


Figure 9.3: Logs and seasonal differences of the A10 (antidiabetic) sales data. The logarithms stabilise the variance, while the seasonal differences remove the seasonality and trend.

[1] `https://otexts.com/fpp3/stationarity.html`

Seasonal ARIMA Models

- in the example above, the differencing was seasonal of order $I = 12$ - i.e. today's values were differenced with last-year's values, on the same day;
- This is called *Seasonal* differencing;
- We can generally extend ARIMA models to include Seasonal components, defined as:
  $SARIMA(p, i, q)(P, I, Q)[s]$
  where the second set of components defines the respective seasonal AR, differencing and MA parts of the modeling framework.

Seasonal ARIMA models

✏ Example: $SARIMA(1,1,2)(1,1,1)[12]$

$$\nabla_{12}\nabla_1 y_t = \phi_1 \nabla_{12}\nabla_1 y_{t-1} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \Phi_1 \nabla_{12}\nabla_1 y_{t-12} + \Theta_1 \epsilon_{t-12} + \epsilon_t$$

- $\nabla_{12}\nabla_1 y_t = \nabla_1 y_t - \nabla_1 y_{t-12} = (y_t - y_{t-1}) - (y_{t-12} - y_{t-12-1})$
- ∇_1 is the regular differencing component;
- ∇_{12} is the seasonal differencing component;
- ϕ is the regular auto-regressive component;
- Φ is the seasonal auto-regressive component;
- θ is the regular moving-average component;
- Θ is the seasonal moving-average component.

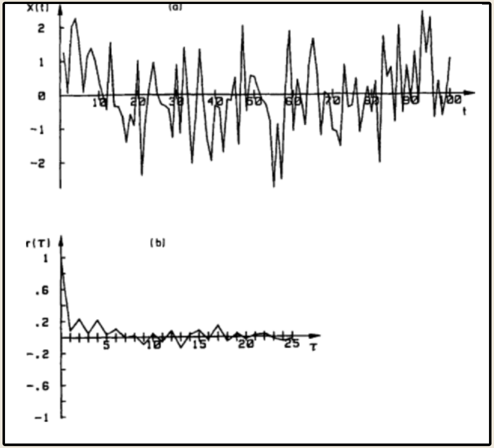
Autocorrelation

- We have described time series models in terms of their Expected Values and Variance...
- but a key feature of time-series is their **autocorrelation**;
- the autocorrelation function (ACF) is the sequence of pearson-correlation values for the current value of the series y_t, at every lag k:

$$R(k) = \frac{\sum_{t=1}^{T-k}(y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

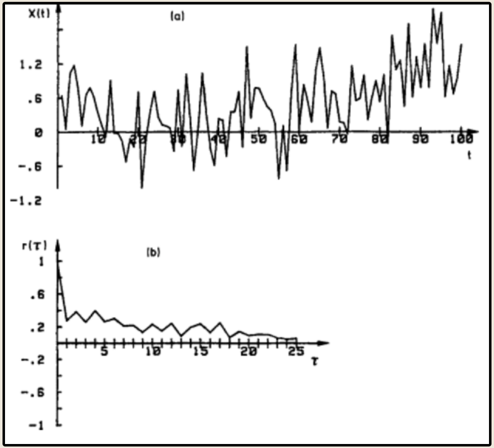
- R is plotted on a *correlogram*;
- features of the ACF of a given series can help us identify which model generated the series...

ACF: AR(1)



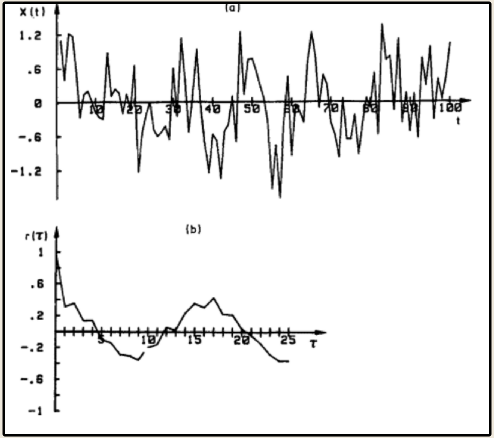
[1]Massart, D. L. (1988). Data handling in science and technology. Chemometrics: a textbook, 2.

ACF: AR(1) + with Drift



[1] Massart, D. L. (1988). Data handling in science and technology. Chemometrics: a textbook, 2.

ACF: Seasonal AR(15)



[1] Massart, D. L. (1988). Data handling in science and technology. Chemometrics: a textbook, 2.

Partial Autocorrelation Function

- ✏ ACF: the total correlation between two points in time. This includes:
  - the direct relationship between those two points;
  - indirect correlations that might be mediated through their relationships with other points in the series.
- ✏ To isolate the direct relationship between two points in time, without the confounding influence of their relationships with intermediate points, we use the Partial Autocorrelation Function (PACF).
- ✏ In crude terms, this is simply the coefficient ϕ_k on lag k on a regression that include AR(K) coefficeints:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} \dots \phi_k y_{t-k}$$

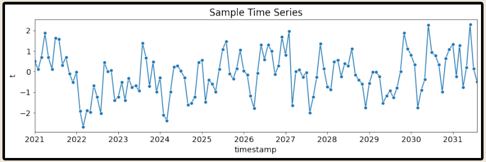
Choosing the Correct Order of a SARIMA(p,i,q)(P,I,Q)[s] Model[2]

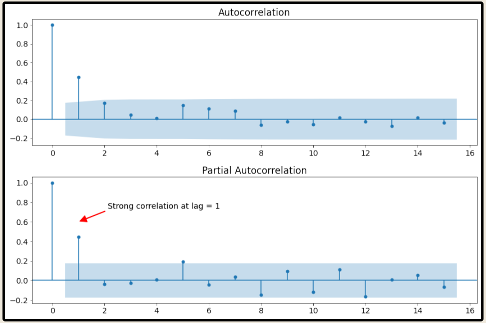
Model	ACF	PACF
AR(p)	Damped exponential and/or sine functions	$\phi_k = 0 \quad \forall k > p$ (Cuts off after lag p)
MA(q)	$R_k = 0 \quad \forall k > q$ (Cuts off after lag q)	Dominated by damped exponential and/or sine functions
ARMA(p,q)	Damped exponential and/or sine functions after lag $\max(0, p-q)$	Dominated by damped exponential and/or sine functions after lag $\max(0, p-q)$

Table: Identifying order of ARIMA models. For seasonal components, the same behaviours appear, and they repeat every s periods.

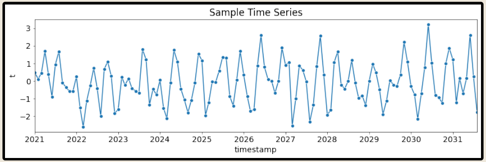
[2] https://www.kaggle.com/code/iamleonie/time-series-interpreting-acf-and-pacf

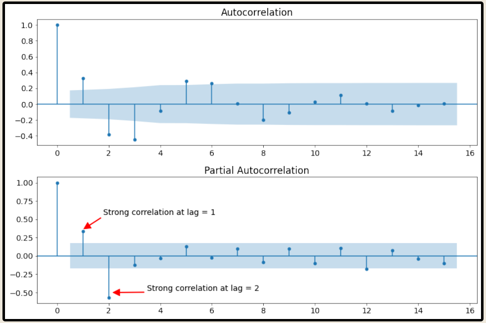
ACF / PACF Examples



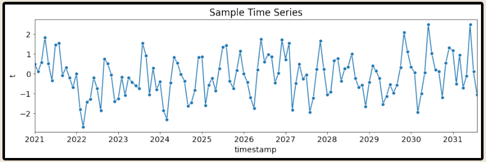


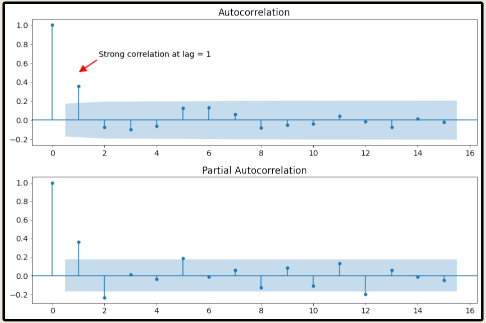
ACF / PACF Examples



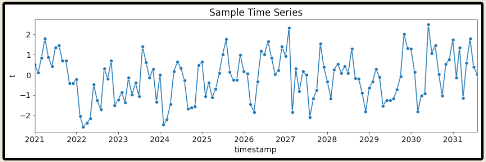


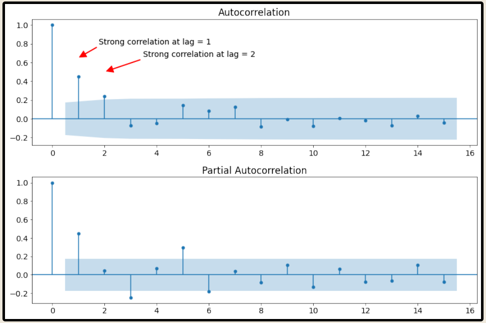
ACF / PACF Examples



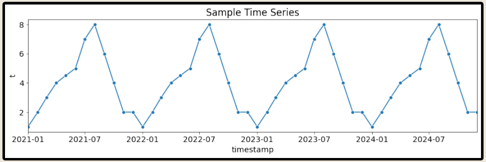


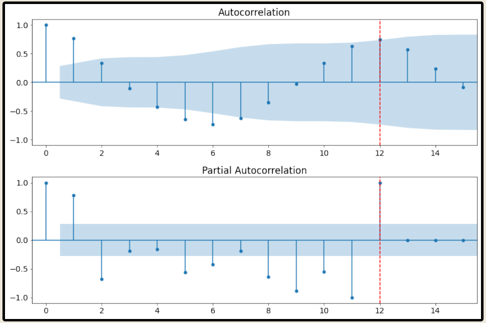
ACF / PACF Examples



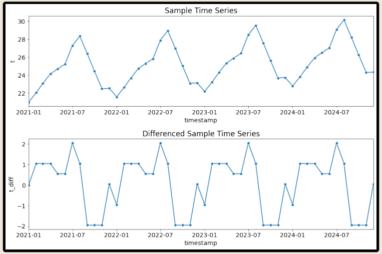


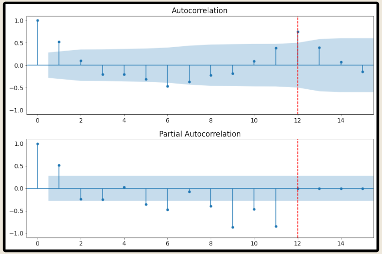
ACF / PACF Examples



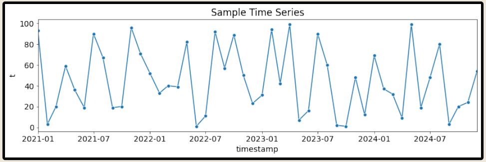


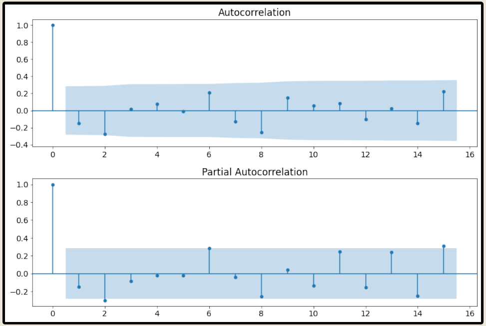
ACF / PACF Examples



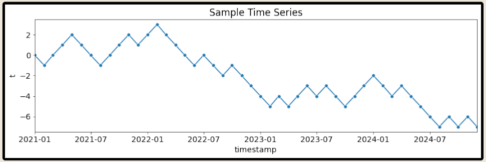


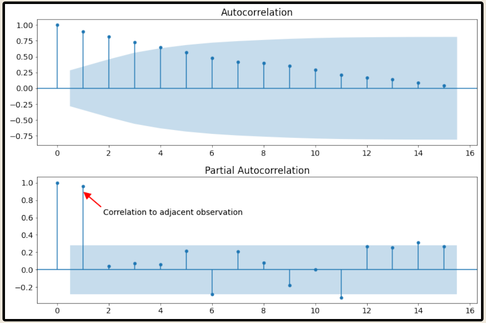
ACF / PACF Examples





ACF / PACF Examples





ACF / PACF Examples