Financial Analytics - conceitos de séries temporais

PADS

Paloma Vaissman Uribe

Insper

Jul 2023

Objetivos da aprendizagem: curso Financial Analytics

Ao final o curso saberemos:

- Analisar dados que variam no tempo, identificando componentes como tendências/ sazonalidades e ruídos, especialmente aplicados em dados de séries financeiras.
- Selecionar métodos estatísticos adequados aos dados e ao problema de decisão;
- Ajustar e interpretar resultados de modelos de séries temporais, em especial, modelos de volatilidade e modelos de regressão CAPM;
- Desenvolver códigos em R ou Python para análise e modelagem de séries de tempo.

Estrutura do curso

1. Conceitos de séries temporais
2. Modelagem SARIMA e Prophet
3. Fatos estilizados em finanças
4. Modelos de volatilidade
5. Aplicações em finanças (VaR, CAPM)
6. Modelos lineares dinâmicos
7. Tópicos especiais

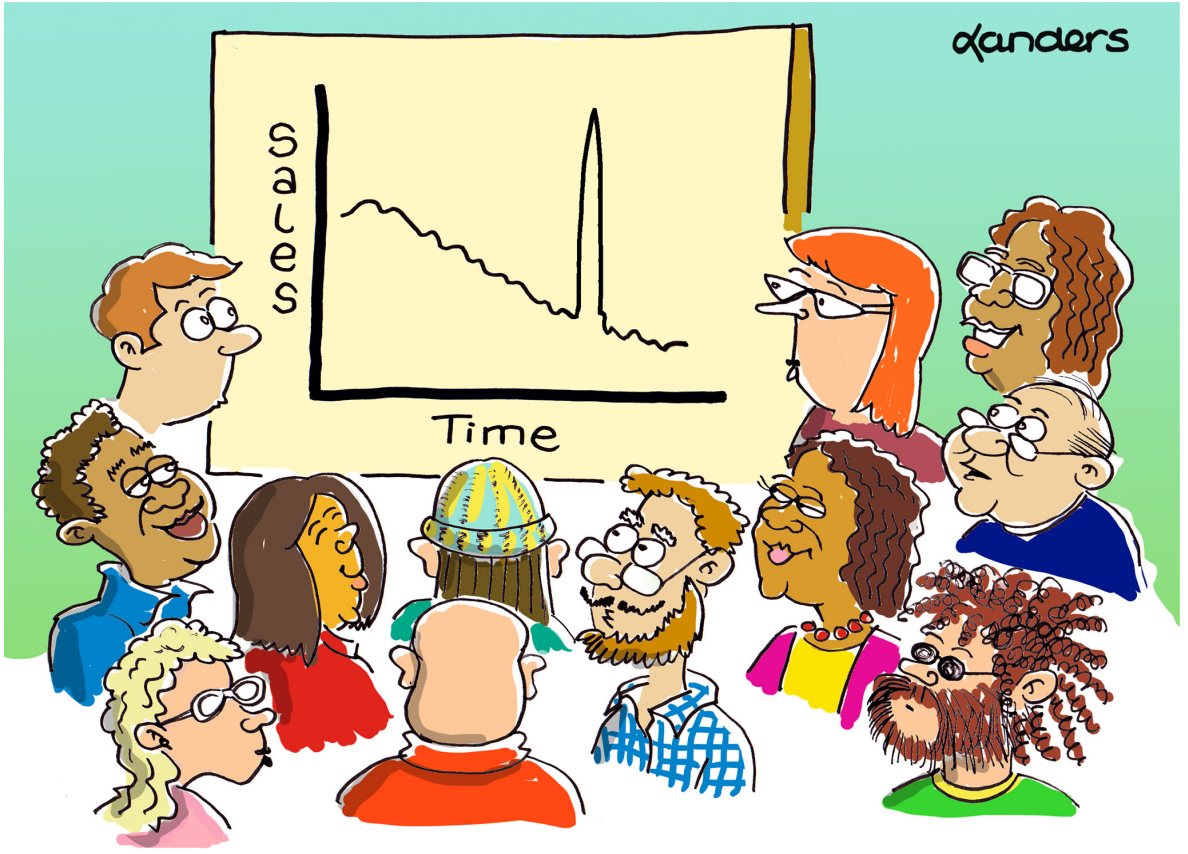
Método de avaliação

- Trabalho intermediário: 40%
  - Somente entrega
- Trabalho em grupo: 60%
  - Apresentação em grupo na última aula

Nesta aula

- Aula 1: conceitos de séries temporais
  - autocorrelação
  - processo estocástico
  - ruído branco
  - componentes: tendência, sazonalidade, ruído
  - tipos de tendência
  - teste ADF
  - modelos ARMA
  - modelos ARIMA e SARIMA
  - Ajuste de modelos ARIMA/SARIMA via metodologia Box-Jenkins.

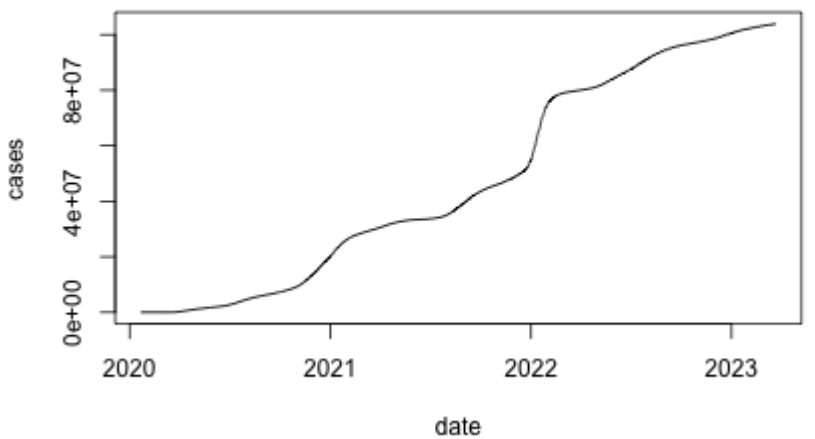
Séries temporais: o estudo de dados ao longo do tempo



Motivação: Covid-19

Vamos importar o arquivo de dados de Covid-19 dos EUA e plotar a série de casos confirmados X_t. Podemos verificar que há uma forte tendência de aumento dos casos ao longo do tempo (com relativa estabilização mais recentemente):

```
library(readr)
us <- read_csv("https://raw.githubusercontent.com/nytimes/covid-19-da
plot(us$date,us$cases,type='l',ylab='cases',xlab='date')
```



Motivação: Covid-19

Intuição: Podemos pensar que o número de casos da doença no dia tem muita relação com o número de casos do dia anterior, já que se trata de uma doença infecciosa de alto contágio, certo?

Vamos então calcular a correlação entre a série de casos confirmados X_t e a série defasada em um período X_{t-1}.

```
library(dplyr)
us <- us %>% mutate(us_lag=lag(cases,1))
cor(us$cases,us$us_lag,use='na.or.complete')
```

```
## [1] 0.9999939
```

Verificamos que a correlação é quase perfeita!!!

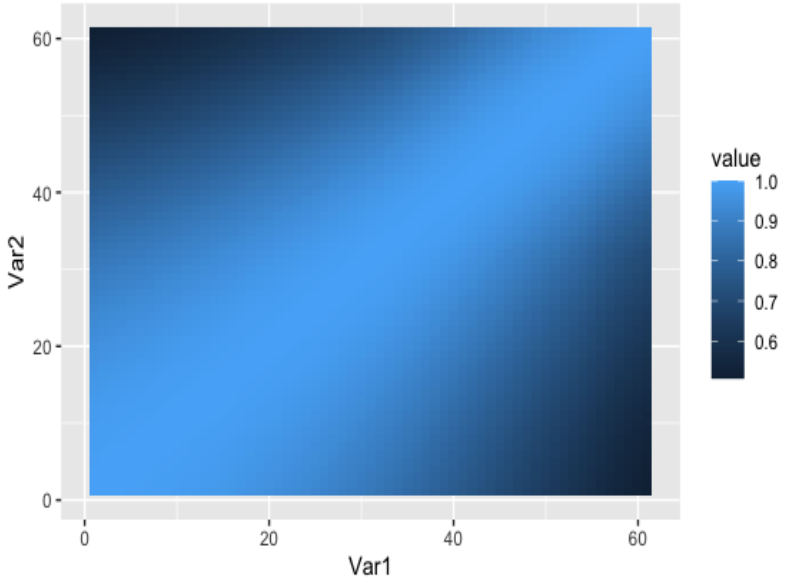
Motivação: Covid-19

Mas o que é **defasagem**?

date	cases	deaths	lag1_cases	lag2_cases
2/29/20	70	1		
3/1/20	88	3	70	
3/2/20	104	6	88	70
3/3/20	125	10	104	88
3/4/20	161	12	125	104
3/5/20	228	12	161	125
3/6/20	311	15	228	161
3/7/20	428	19	311	228
3/8/20	547	22	428	311
3/9/20	748	26	547	428
3/10/20	1018	31	748	547
3/11/20	1263	37	1018	748
3/12/20	1668	43	1263	1018

Motivação: Covid-19

E se calcularmos as correlações das séries defasadas $X_{t-2}, X_{t-3}, \ldots, X_{t-60}$?



Podemos ver que no máximo, as correlações das defasagens com a série original chegam a 50%, mesmo considerando 60 dias de defasagem.

Quando isso acontece, dizemos que a série possui "memória longa".

Explicação sobre memória longa

A **memória longa** ou **dependência temporal** é um fenômeno próprio de dados que variam no tempo, i.e., medidas sequenciais que são observadas e guardadas ao longo de um período de tempo (e geralmente em instantes equidistantes): Por exemplo:

- Temperaturas máximas e mínimas diária em uma região da cidade medida ao longo dos meses;
- Preço de fechamento diário de uma ação;
- Vendas totais diárias de uma loja.

A dependência temporal decorre do fato de que essas medidas são na verdade observações de um processo estocástico que não é independente, de tal forma que o dado de hoje tem relação com o dado de instantes anteriores.

Breve revisão estatística: variável aleatória

- **Variável aleatória**: é uma função matemática que associa a cada elemento $\omega \in \Omega$ pertencente ao **espaço amostral** Ω um único número real, isto é, $X : \Omega \to \mathbb{R}$.
- Exemplo: em um evento aleatório de jogar os dados, podemos definir o espaço amostral (todos os possíveis resultados) como sendo $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- Assim, definimos o resultado de jogar o dado não viciado uma vez como sendo uma variável aleatória X, tal que a probabilidade de $X = 1$ é igual a $P(X = 1) = 1/6$.

Breve revisão estatística: processo i.i.d.

- Um **processo estocástico** é uma sequência de variáveis aleatórias.
- Quando essas VAs são independentes e possuem a mesma **distribuição de probabilidade** dizemos que é um processo i.i.d. (**independente e identicamente distribuído**).
- Por exemplo, podemos pensar em jogar um dado sequencialmente, por 500 vezes, e anotar o resultado consecutivo dessas rodadas:

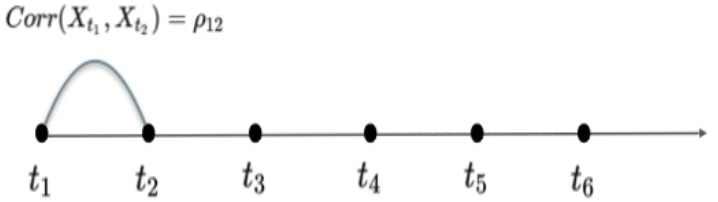
```
n <- 500
dados <- function(n,k){
   sample(1:k,n,replace=T)
}
ts.plot(dados(n,6))
```

- Contudo, na análise de séries temporais, a grande diferença é que não se trata, em geral, de um processo i.i.d., pois há o que chamamos de **dependência temporal** ou **memória**.

Problemas novos e únicos na modelagem estatística

A **correlação entre pontos adjacentes no tempo** restringe a aplicabilidade de métodos estatísticos convencionais que dependem do pressuposto de que as observações são independentes e identicamente distribuídas (i.i.d.).

O que de fato é autocorrelação?



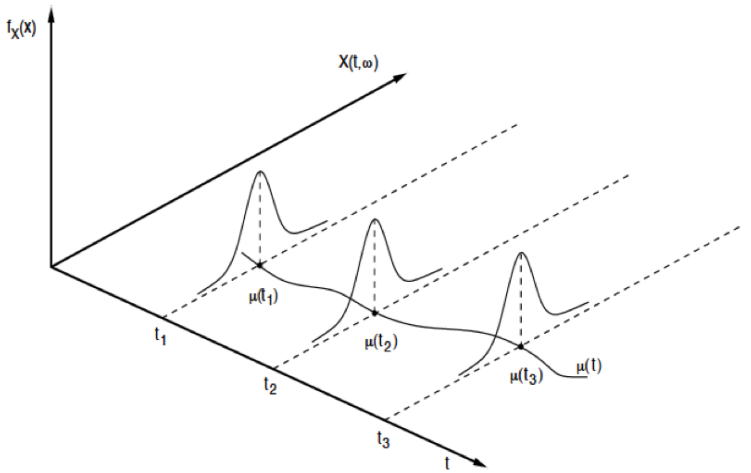
Em geral, espera-se que à medida que aumenta-se a defasagem a (auto) correlação ou memória da série caia - **dizemos que a função de autocorrelação da série X_t decai para zero**, ou seja,
$\rho_{16} < \rho_{15} < \rho_{14} < \rho_{13} < \rho_{12} < \rho_{11}$.

Processo estocástico como uma família de variáveis aleatórias

Um **processo estocástico** é definido por

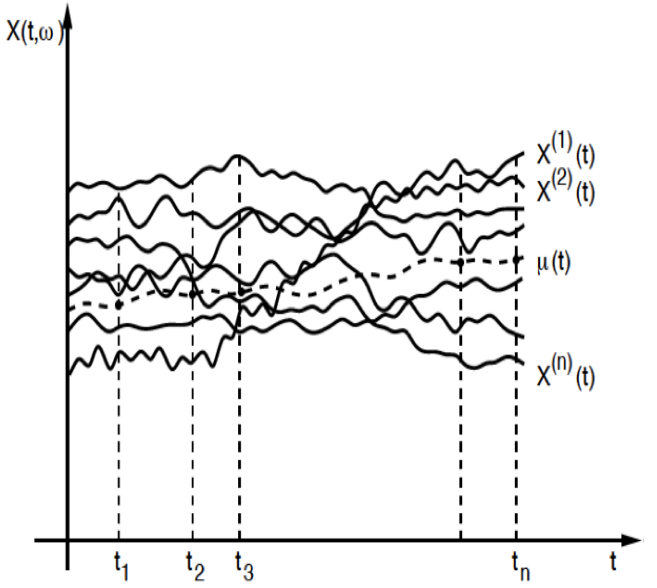
$$X = \{X(t, w), t \in T, w \in \Omega\}$$

Considerando t **fixado**, $X(t, w) = X(w)$ é uma variável aleatória definida sobre o espaço amostral Ω, e para cada w **fixado**, $X(t, w) = X(t)$ é uma função de t, denominada realização (trajetória) de um processo estocástico, ou simplesmente, **série temporal**.



Processo estocástico como uma família de trajetórias

Vamos designar as realizações de um processo estocástico por $X^{(1)}$, $X^{(2)}$, etc. Cada realização é uma função de t, e para cada t fixo, $X(t)$ é um número real ou complexo.



Ou seja, existem várias trajetórias (séries temporais observadas) possíveis para um processo estocástico.

Processo estacionário

Dizemos que um processo é estacionário quando:

- **Visualmente**: percebe-se que há uma **reversão à média**, i.e., o gráfico da série mostra que esta oscila em torno de uma média constante;
- **Formalmente**: Quando analisamos a **função de autocorrelação** de uma série, estamos analisando qual a correlação entre as variáveis aleatórias entre dois pontos do tempo.

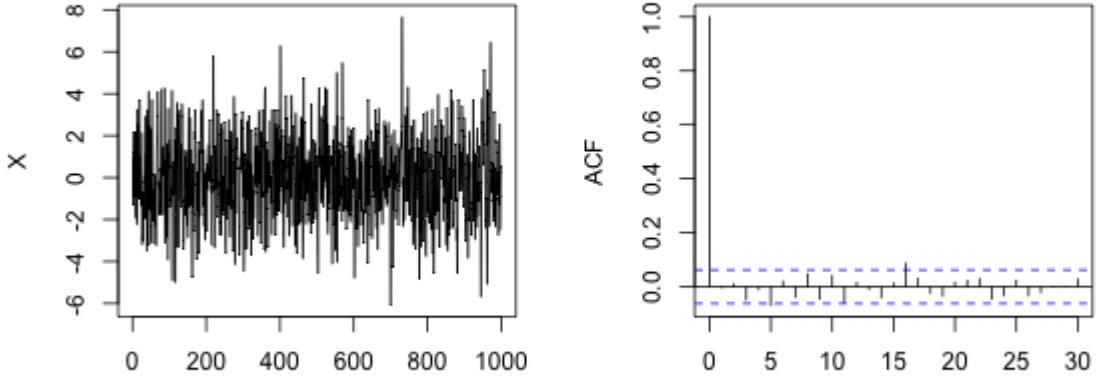
Um processo **fracamente estacionário** (ou simplesmente estacionário) caracteriza-se por ter:

- *média* e a *variância* constantes para todo $t \in T$ e
- *covariância* $Cov(X(t_i), X(t_j)) = \rho_{ij}$ sendo uma função apenas da defasagem, ou intervalos entre instantes de tempo.
- Na prática: vemos que a covariância e a correlação entre $X(t_1)$ e $X(t_2)$ é igual à correlação entre $X(t_2)$ e $X(t_3)$ e assim por diante...

Função de autocorrelação (FAC ou ACF) de um processo estacionário

Podemos plotar a função de autocorrelação (amostral) de uma série no R. Para isso, vamos começar com um processo i.i.d (que como vimos não possui dependência temporal, visto que as V.A.s são independentes). Usaremos a função básica do R `acf` (autocorrelation function):

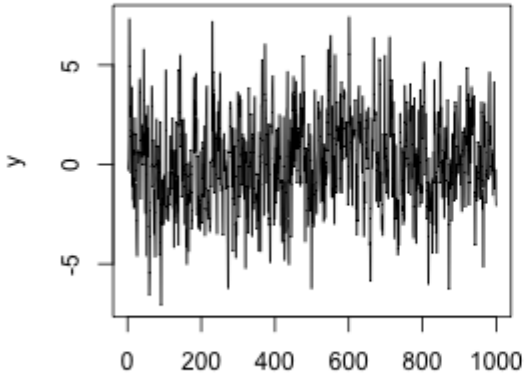
```
X <- rnorm(1000,mean=0,sd=2)
par(mfrow=c(1,2))
ts.plot(X)
acf(X,main="")
```

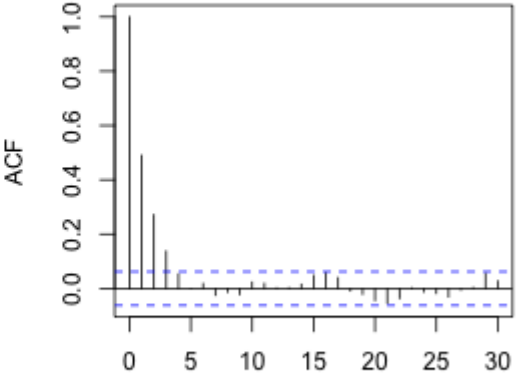


Voltando ao processo $y_t = \phi y_{t-1} + \varepsilon_t$

Nesse caso a FAC não é sempre igual à zero para lags diferentes de zero, porém decai rapidamente.

```
phi <- 0.5
n <- 1000
e <- rnorm(n,mean=0,sd=2)
y <- rep(NA,n)
y[1] <- e[1]
for (i in 2:n){
  y[i] <- phi*y[i-1] + e[i]
}
```

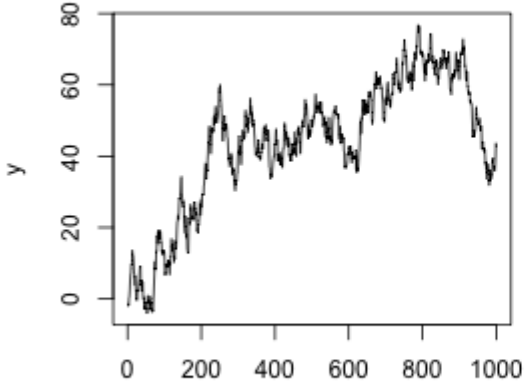


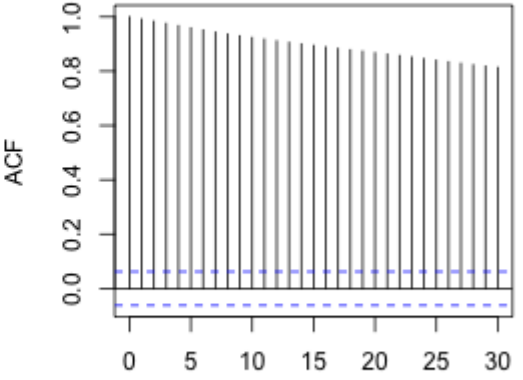


E se $\phi = 1$?

Nesse caso a FAC não porém decai rapidamente. Veremos que esse processo possui um nome específico (aka Passeio Aleatório)

```
phi <- 1
n <- 1000
e <- rnorm(n,mean=0,sd=2)
y <- rep(NA,n)
y[1] <- e[1]
for (i in 2:n){
  y[i] <- phi*y[i-1] + e[i]
}
```





Passeio aleatório

Um processo é um **passeio aleatório** (random walk) é definido por

$$y_t = y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim RB(0, \sigma^2).$$

Na literatura de séries temporais, este processo é chamado também um processo de tendência estocástica ou que possui raíz unitária .

Ruído Branco

- Estatisticamente, dizemos que $\{\varepsilon_t, t \in Z\}$ é um **ruído branco** se a covariância para as defasagens diferentes de zero *são sempre iguais à zero*.

$$Cov(\varepsilon_t, \varepsilon_s) = 0, \forall t \neq s.$$

- Isso equivale a dizer que trata-se de um processo i.i.d!!!
- Vocês verão que é bem comum definir especificações em séries temporais em termos de um erro que é um ruído branco.

Análise gráfica de séries de tempo

- Até aqui vimos alguns conceitos importantes relacionados à análise de séries temporais:
    - Ruído Branco
    - Estacionariedade de um processo
    - Passeio aleatório
- Agora vamos ver outro conceito interessante que auxilia na análise gráfica das séries e nos dá uma noção de qual especificação (modelo) utilizar: **os componentes de uma série**.

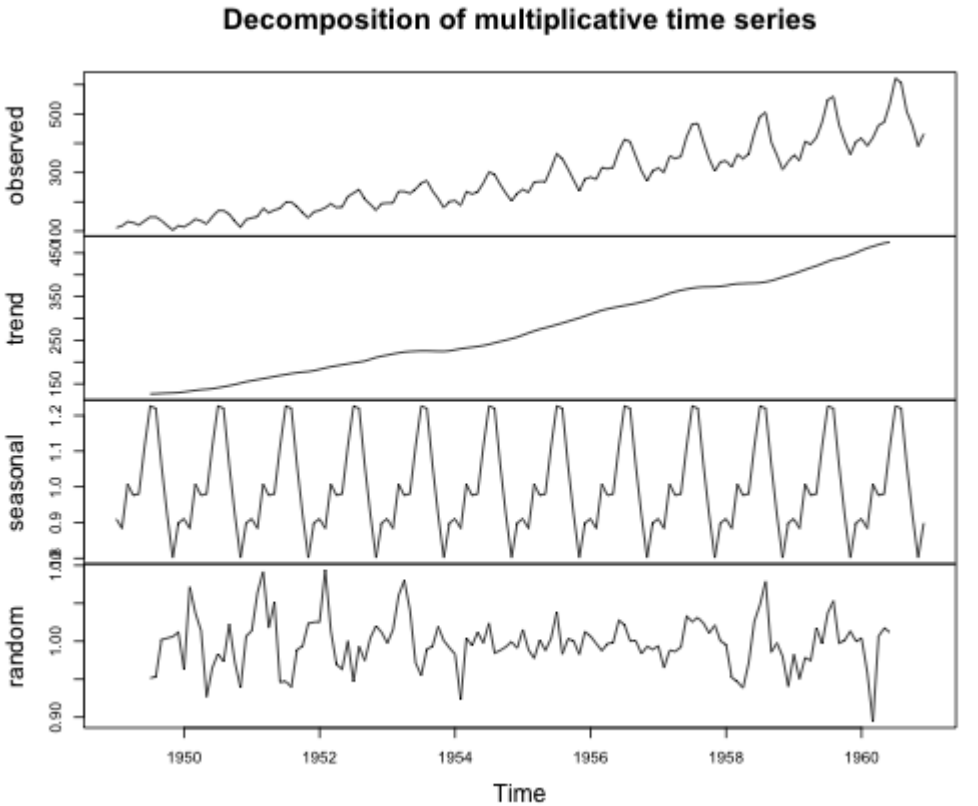
Componentes de uma série temporal

- **Nível** L_t é um componente sistemático que representa o valor médio da série, podendo variar no tempo em certos casos.
- **Tendência** T_t é um um componente linear sistemático ou (na maioria das vezes) não-linear que muda com o tempo e não se repete, representando variações de baixa frequência, frequentemente associado a um movimento de longo prazo.
- **Sazonalidade** S_t é um componente linear sistemático geral ou (na maioria das vezes) não linear que muda com o tempo e se repete.
- **Ruído** R_t é um componente não sistemático que não se repete e não tem padrão, representando variações aleatórias.

Componentes de uma série I

Podemos pensar em **modelo multiplicativo**:

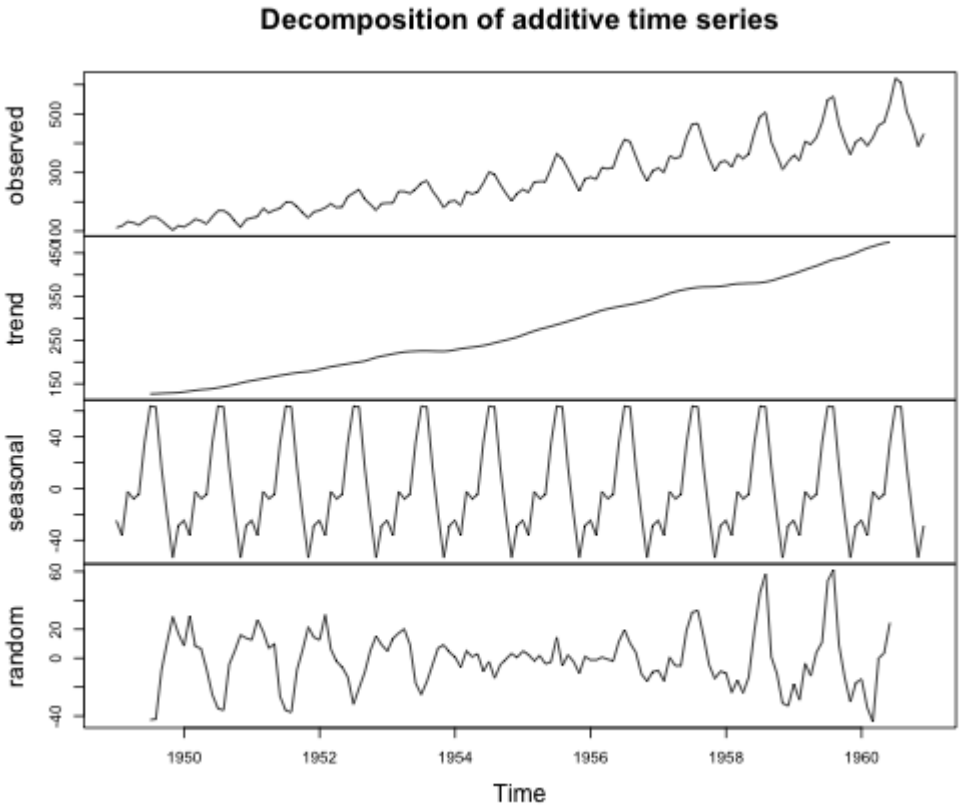
$$Y_t = T_t * S_t * R_t$$



Componentes de uma série II

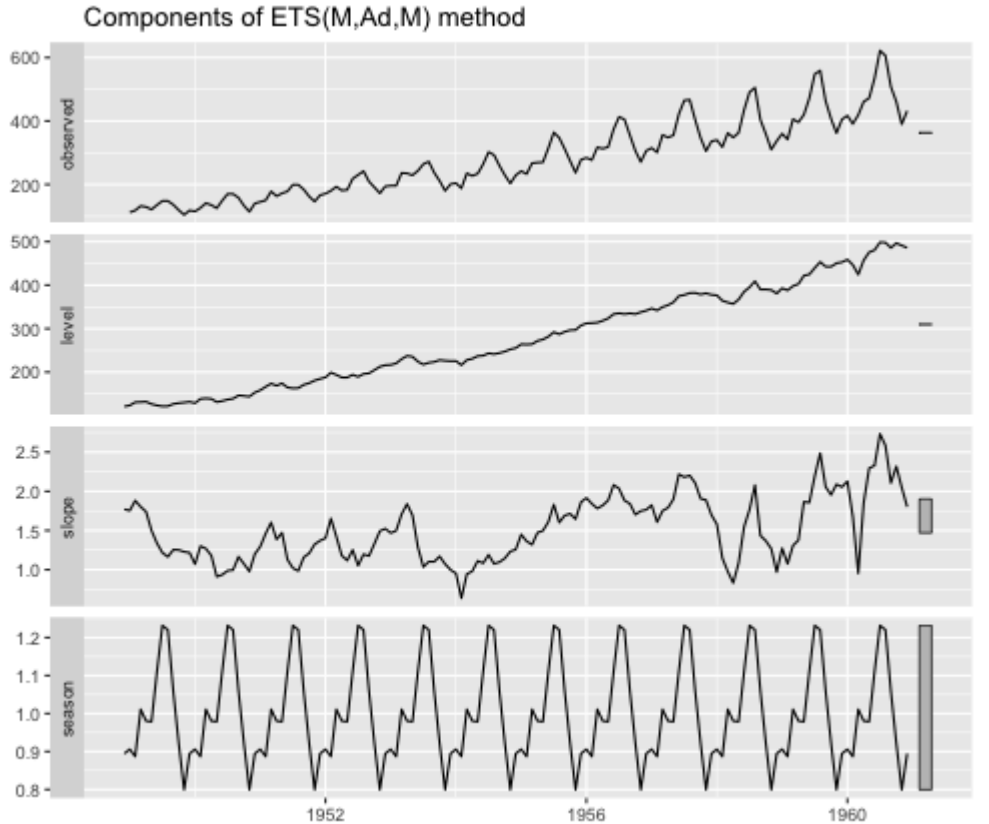
Ou então em um **modelo aditivo**:

$$Y_t = T_t + S_t + R_t$$



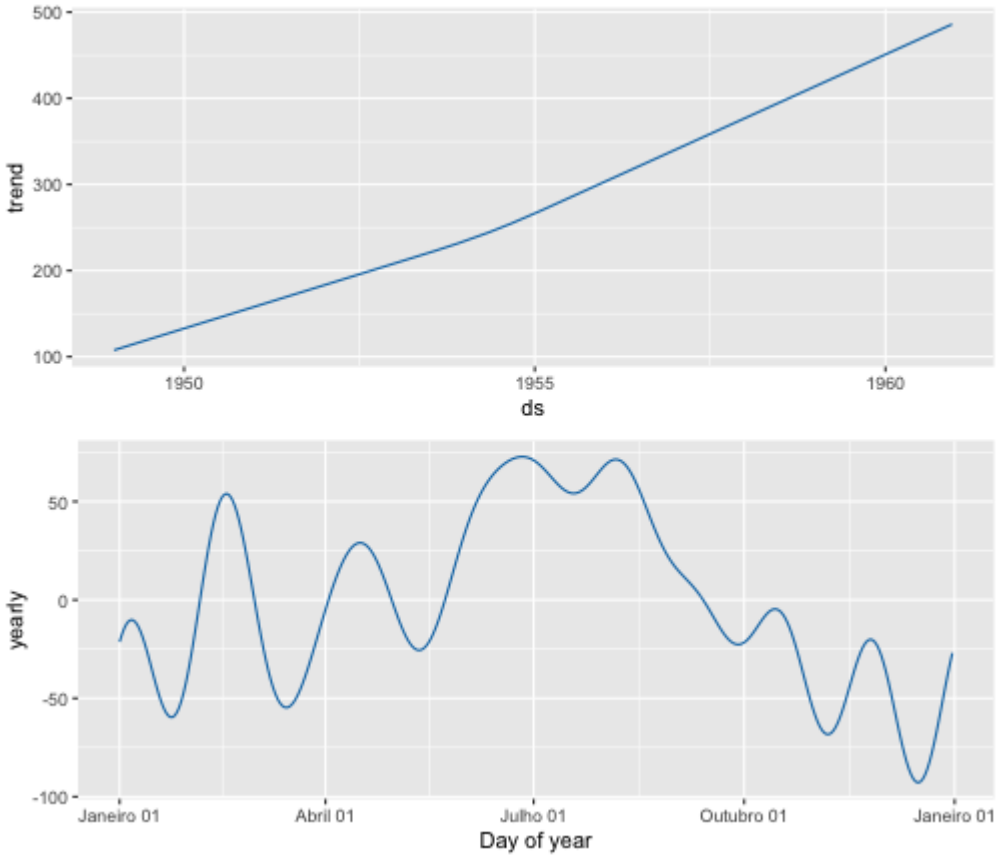
Outros pacotes de decomposição

Pacote `forecast` tem a decomposição ETS (*Error, Trend, Seasonality*):



Outros pacotes de decomposição

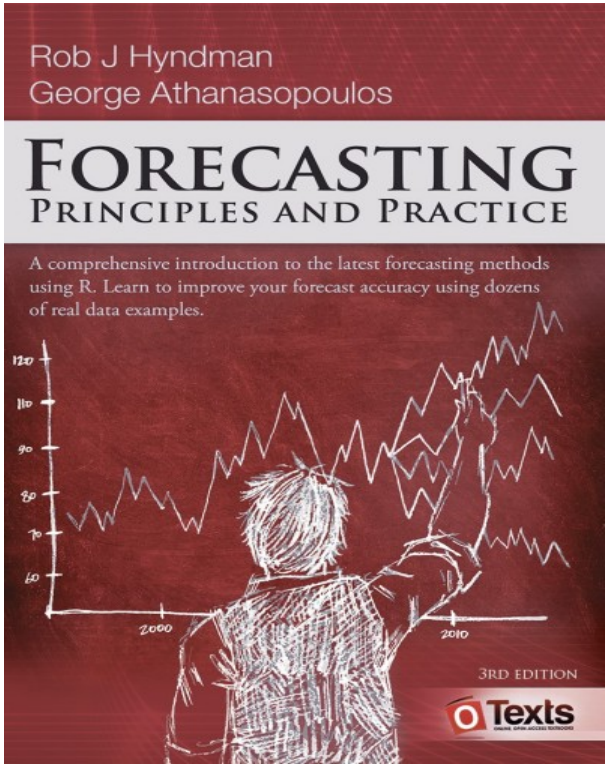
E também o pacote `prophet`:



Livro recomendado

Uma ótima referência sobre o tema de decomposição de séries de tempo é o site e livro

- *Forecasting: Principles and Practice (3rd ed)*. **Rob J Hyndman and George Athanasopoulos**. Monash University, Australia, disponível em: https://otexts.com/fpp3/



Exercício

Vamos trabalhar com a série **us_retail_employment** do pacote `fpp3`.

1. Plotar o gráfico da série
2. Verificar a ACF e tecer comentários
3. Fazer decomposição (ETS, STL, R basis, Prophet)
4. Dessazonalizar a série

Nota sobre a tendência de uma série

- Vimos que a tendência é um componente importante de uma série. Em muitos casos, os processos que tem tendência são não estacionários e exibem o que chamamos de **raíz unitária** ou **tendência estocástica** ou **passeios aleatórios**.
- **Como saber se essa tendência é desse tipo?**
  - Existem testes estatísticos para detectar isso, o mais comum é o **teste ADF** (Augmented Dicker-Fuller). Veremos como realizar esse teste usando a série GDP-EUA disponível no BlackBoard (fonte: site FRED).
  - Usaremos o pacote `tseries` e a recomendação de uso na documentação: https://cran.r-project.org/web/packages/tseries/tseries.pdf

```
library(tseries)
library(readr)
GDP_EUA <- read_csv("GDP_EUA.csv")
gdp = GDP_EUA$NA000334Q
adf.test(gdp, alternative = c("stationary", "explosive"),
k = trunc((length(gdp)-1)^(1/3)))
```

Resultado teste ADF

```
##
##     Augmented Dickey-Fuller Test
##
## data:  gdp
## Dickey-Fuller = 1.2887, Lag order = 6, p-value = 0.99
## alternative hypothesis: stationary
```

A hipótese nula é que existe raíz unitária, logo não rejeitamos essa hipótese, dado o alto p-valor.

Exercício

- Realizar o teste de raíz unitária na série **us_retail_employment** do pacote `fpp3`. Note que o pacote em questão não posssui o teste ADF, mas tem o teste KPSS (Kwiatkowski-Phillips-Schmidt-Shin), podendo optar por este, caso queira (nesse caso a hipótese nula é de estacionariedade).

```
## # A tibble: 148 × 3
##    Series_ID     kpss_stat kpss_pvalue
##    <chr>             <dbl>       <dbl>
##  1 CEU0500000001    12.2          0.01
##  2 CEU0600000001     4.97         0.01
##  3 CEU0800000001    12.1          0.01
##  4 CEU1000000001     2.83         0.01
##  5 CEU1011330001     7.62         0.01
##  6 CEU1021000001     0.946        0.01
##  7 CEU1021100001     1.82         0.01
##  8 CEU1021200001     4.59         0.01
##  9 CEU1021210001     5.30         0.01
## 10 CEU1021300001     4.80         0.01
## # … with 138 more rows
```