Modelo Auto-regressivo Integrado de Médias Móveis para a temperatura média anual nos anos de 1869-2019 do Central Park

Ana Zucon;
Vinicius Riffel

UFPR

Mar. 15, 2021

1. Introdução
2. Materiais e métodos
3. Aplicação em dados reais
4. Conclusão
5. Referências

Introdução

- Os modelos para séries temporais são úteis em situções que temos dependência temporal na variável resposta.

Introdução

- Os modelos para séries temporais são úteis em situções que temos dependência temporal na variável resposta.
- A dependência pode ser por elemento intrínseco à variável resposta ou problema na coleta de amostras.

Introdução

- Os modelos para séries temporais são úteis em situções que temos dependência temporal na variável resposta.
- A dependência pode ser por elemento intrínseco à variável resposta ou problema na coleta de amostras.
- Em qualquer dessas situações temos *erros autocorrelacionados*.

Introdução

- Os modelos para séries temporais são úteis em situções que temos dependência temporal na variável resposta.
- A dependência pode ser por elemento intrínseco à variável resposta ou problema na coleta de amostras.
- Em qualquer dessas situações temos *erros autocorrelacionados*.
- Nessas situações, devemos considerar um modelo que seja capaz de incluir essa autocorrelação.

Introdução

Modelos para séries temporais

- Auto-regressivos (AR).
- Médias móveis (MA).
- Auto-regressivos de médias móveis ($ARMA$).
- Auto-regressivos integrados de médias móveis ($ARIMA$).

Materiais e métodos

Modelos *AR*

- O modelo *AR* de ordem p, denotado por $AR(p)$, é definido por:

$$\hat{Y}_t = c + \sum_{t=1}^{p} \phi_i Y_{t-i} + \varepsilon_t \quad t = p+1, p+2, \ldots, n \tag{1}$$

em que Y_t é um processo estocástico estacionário, ε_t é um termo de erro, c e ϕ_i são os parâmetros do modelo.

- A função de autocovariância do modelo AR(1) é dada por:

$$K_y(t, t+\tau) = Cov(Y_t, Y_t + \tau) = \frac{\sigma_\varepsilon^2}{1-\phi^2}\phi^{|\tau|}$$

onde σ_ε^2 é a variância do termo de erro (constante).

Materiais e métodos

Modelos *MA*

- A equação do modelo $MA(q)$ é dada por:

$$\hat{Y}_t = \mu + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \varepsilon_t \quad t = q+1, q+2, \ldots, n \tag{2}$$

em que ϵ_t é o erro no tempo t, ε_t é um termo de erro, μ e θ_i são os parâmetros do modelo.

Materiais e métodos

Modelos *ARMA*

Combinando as equações 1 e 2 temos o modelo $ARMA(p, q)$:

$$\hat{Y}_t = c + \sum_{t=1}^{p} \phi_i Y_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \varepsilon_t$$
$$t = min(p, q) + 1, min(p, q) + 2, \dots, n.$$

Materiais e métodos

Operador de defasagem (LAG)

- Indica quantidade de tempos anteriores em relação ao tempo t:

$$L^n y_t = y_{t-n}, \quad \text{em que } t > n.$$

- Diferenciando a série uma vez:

$$y_t' = y_t - y_{t-1} = y_t - Ly_t = (1 - L)y_t.$$

- Generalizando: $(1 - L)^n y_t = y_t - \sum_{i=1}^{n} y_{t-i}$.

Materiais e métodos

Modelos *ARIMA*

- Usado quando o processo Y_t é não estacionário.
- E equação do modelo $ARIMA(p, d, q)$ é dada por:

$$(1 - L)^d \hat{Y}_t = c + \sum_{i=1}^{p} \phi_i (1 - L)^d Y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \varepsilon_t,$$

$$t < d, d \leq p \text{ e } d \leq q.$$

- p, d e q são a *ordem* do modelo.
- Os parâmetros são c, ϕ_i e θ_i.

Materiais e métodos

Caso especiais de *ARIMA*

- ARIMA(0, 0, 0): Ruído branco (assumindo que $\varepsilon_t \sim N(0, \sigma^2_\varepsilon)$).
- ARIMA(0, 0, q): $MA(q)$.
- ARIMA(p, 0, 0): $AR(p)$.
- ARIMA(p, 0, q): $ARMA(p, q)$.

Materiais e métodos

Estimação dos modelos

- Os estimadores de mínimos quadrados podem ser obtidos por:

$$\underset{\hat{\Upsilon}\in\Theta}{\operatorname{argmin}}\, S(\hat{\Upsilon}) = \sum_{t\in T} \varepsilon_t^2 \tag{3}$$

Ou assumindo que $\varepsilon_t \sim NID(0; \sigma_\varepsilon^2)$:

$$\underset{\hat{\Upsilon}\in\Theta}{\operatorname{argmax}}\, L(\hat{\Upsilon}) = (2\pi)^{-n/2}(\sigma_\varepsilon)^{-n} exp\left(-\sum_{t\in T}\frac{{\varepsilon_t}^2}{2\sigma_\varepsilon^2}\right) \tag{4}$$

em que $\hat{\Upsilon}$ é o vetor de parâmetros do modelo.

- Os estimadores são obtidos minimizando a função 3 ou maximizando a função 4.
- Os estimadores de máxima verossimilhança e mínimos quadrados são idênticos.

Materiais e métodos

IC e TH

- Pode ser provado que:

$$\sqrt{n}(\hat{\Upsilon} - \Upsilon) \xrightarrow{D} N(\mathbf{0}_{r\times 1}; \boldsymbol{I_F}^{-1})$$

em que r é o rank da matriz do modelo e $\boldsymbol{I_F}$ é a matriz de informação de Fisher.

Materiais e métodos

IC e TH

- IC com $(1-\alpha)\%$ de confiança para algum parâmetro υ do modelo é dado por:

$$\hat{\upsilon} \pm t_\alpha \sqrt{\hat{Var}(\hat{\upsilon})}$$

em que t_α é o quantil da distribuição T que fornece $(1-\alpha)\%$ de confiança. Caso σ_ε^2 seja conhecida, o quantil utilizado será da distribuição Normal.

- Da mesma forma, pode-se obter o IC para predição ou para os valores ajustados:

$$\hat{y}_t \pm t_\alpha \sqrt{\hat{Var}(\hat{y}_t)}$$

Seleção de modelos

- Como selecionar um dentre os diferentes modelos?
- Podemos utilizar o Akaike Information Criteria (AIC):

$$AIC = -2l(\hat{\Upsilon}; \boldsymbol{y}) + 2p$$

onde p é o número de parâmetros do modelo e $l(\hat{\Upsilon}; \boldsymbol{y})$ é a moda da função log-verossimilhança.

- Na prática, ajustamos vários modelos $ARIMA$ de diferentes ordens e selecionamos aquele que tem o menor AIC.

Aplicação em dados reais

- Mudanças climáticas é assunto recorrente na comunidade científica.

Aplicação em dados reais

- Mudanças climáticas é assunto recorrente na comunidade científica.
- Vamos modelar as temperaturas médias anuais do Central Park (Nova York) utilizando um modelo ARIMA.

Aplicação em dados reais

- Mudanças climáticas é assunto recorrente na comunidade científica.
- Vamos modelar as temperaturas médias anuais do Central Park (Nova York) utilizando um modelo ARIMA.
- Os dados foram coletados de 1869 até 2019.

Aplicação em dados reais

- Mudanças climáticas é assunto recorrente na comunidade científica.
- Vamos modelar as temperaturas médias anuais do Central Park (Nova York) utilizando um modelo ARIMA.
- Os dados foram coletados de 1869 até 2019.
- As medidas estão em graus Fahrenheit.

Aplicação em dados reais

- Mudanças climáticas é assunto recorrente na comunidade científica.
- Vamos modelar as temperaturas médias anuais do Central Park (Nova York) utilizando um modelo ARIMA.
- Os dados foram coletados de 1869 até 2019.
- As medidas estão em graus Fahrenheit.
- A análise foi construída no software `R`, através do pacote `forecast`.

Aplicação em dados reais

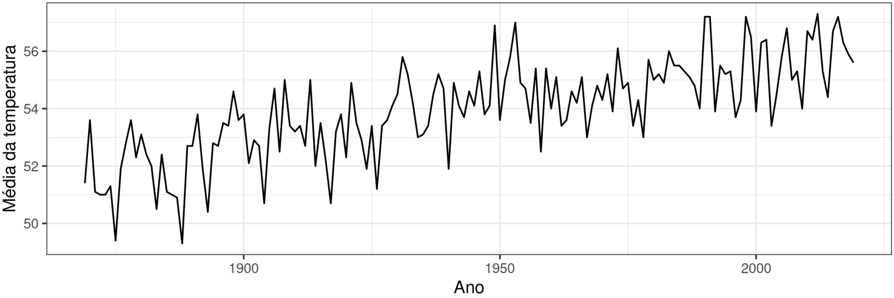

Figura 1: Gráfico da série com $LAG = 0$.

Aplicação em dados reais

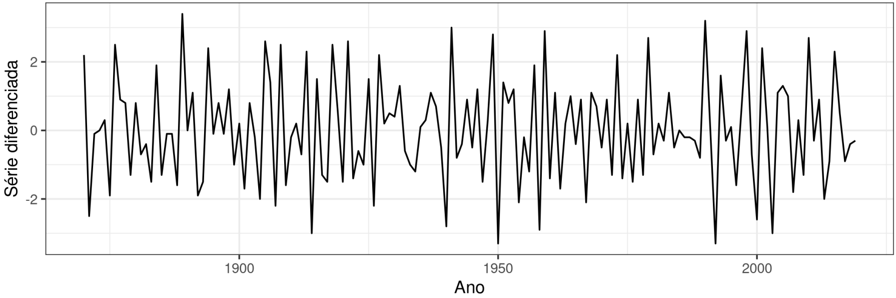

Figura 2: Gráfico da série com $LAG = 1$.

Aplicação em dados reais

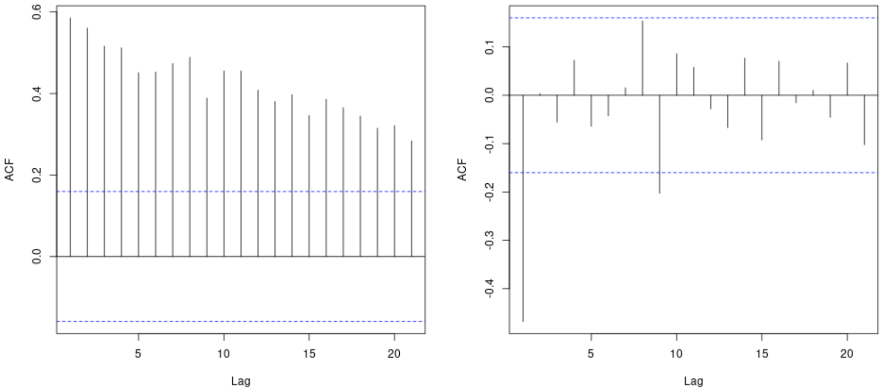

Figura 3: Gráfico das autocorrelações da série com $LAG = 0$ (esquerda) e $LAG = 1$ (direita).

Aplicação em dados reais

Modelo	**AIC**	$\hat{c}$	$\hat{\theta}$	$\hat{\phi}$
ARIMA(0,0,0)	590.84	53.97		
ARIMA(0,0,1)	556.85	53.96	0.39	
ARIMA(0,1,0)	553.88			
ARIMA(1,0,0)	528.15	53.96		0.59
ARIMA(0,1,1)	**481.78**		**-0.86**	
ARIMA(1,0,1)	489.63	53.87	-0.85	0.99
ARIMA(1,1,0)	518.51			-0.47
ARIMA(1,1,1)	483.28		-0.87	0.06

Aplicação em dados reais

```r
(fit <- Arima(serie, order = c(0, 1, 1), method = "ML"))
```

```
## Series: serie
## ARIMA(0,1,1)
##
## Coefficients:
##           ma1
##       -0.8561
## s.e.   0.0406
##
## sigma^2 estimated as 1.412:  log likelihood=-238.89
## AIC=481.78   AICc=481.86   BIC=487.8
```

Aplicação em dados reais

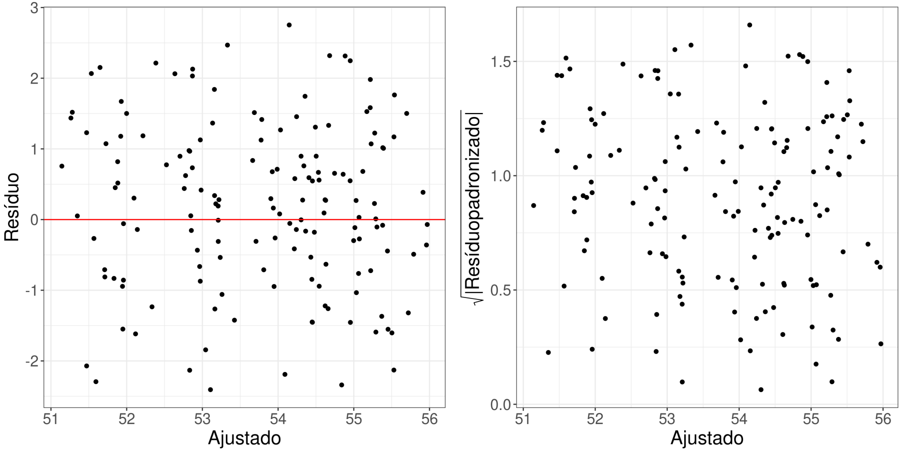

Figura 4: Resíduos ordinários (esquerda) e raíz quadrada do valor absoluto dos resíduos padronizados (direita) para cada valor ajustado.

Aplicação em dados reais

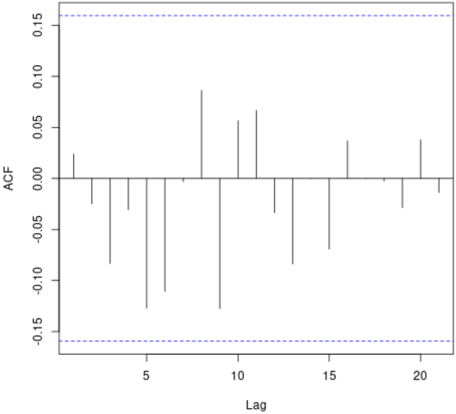

Figura 5: Gráfico ACF dos resíduos.

Aplicação em dados reais

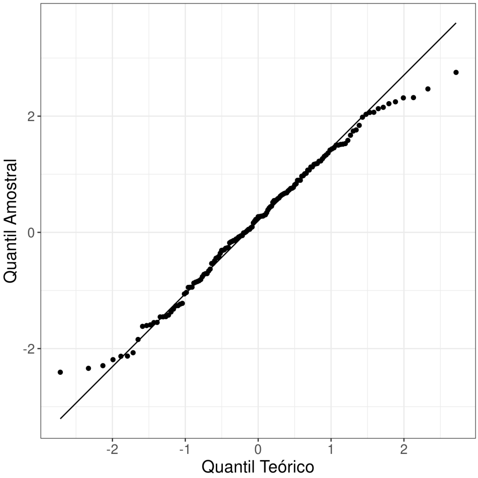

Figura 6: Gráfico quantil-quantil dos resíduos.

Aplicação em dados reais

- A equação do modelo ajustado é dada por:

$$\hat{Y}_t = Y_{t-1} - 0.8561(\hat{Y}_{t-1} - Y_{t-1}) \text{ com } t = 2, 3, ..., 151.$$

- Para esse modelo, a variância de $\hat{\theta}$ é estimada por:

$$\hat{Var}(\hat{\theta}) \approx \frac{1 - (-0.8561)^2}{150} \approx 0.0016$$

- Para $\alpha = 0.05$, o intervalo de confiança/predição é dado por:

$$\hat{y}_t \pm t_{0.05}\sqrt{1.412(1 + \tau(1 - 0.8561)^2)}$$

em que τ é a quantidade de tempos adiante de t que iremos prever.

Aplicação em dados reais

- Para testar se o coeficiente é signficativo, fazemos:

$$t_{obs} = \frac{-0.8561}{0.04} \approx -21.$$

- O intervalo com 95% intervalo de confiança para θ pode ser obtido a partir da função `confint`.

```
confint(fit)
```

```
##          2.5 %     97.5 %
## ma1 -0.9356608 -0.7766035
```

Aplicação em dados reais

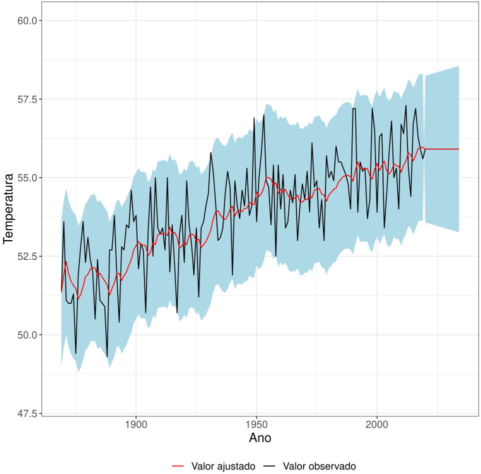

Figura 7: Gráfico com intervalo de confiança dos valores ajustados e um intervalo de predição até 2024.

Conclusão

- Séries temporais são técnicas estatísticas que podem ser utilizadas nas mais diferentes áreas de pesquisa.

Conclusão

- Séries temporais são técnicas estatísticas que podem ser utilizadas nas mais diferentes áreas de pesquisa.
- Apresentamos diversos modelos para dados com dependência temporal.

Conclusão

- Séries temporais são técnicas estatísticas que podem ser utilizadas nas mais diferentes áreas de pesquisa.
- Apresentamos diversos modelos para dados com dependência temporal.
- Outro tipos de modelos para séries temporais lidam com sazonalidade e situações em que nossos dados são vetores (modelos multivariadas).

Referências

J. Hamilton. **Time Series Analysis**. Princeton University Press. 1994.

R. Hyndman, G. Athanasopoulos, C. Bergmeir, G. Caceres, L. Chhay, M. O'Hara-Wild, F. Petropoulos, S. Razbash, E. Wang, and F. Yasmeen. **forecast: Forecasting functions for time series and linear models**, 2020. R package version 8.13.

R. J. Hyndman and Y. Khandakar. **Automatic time series forecasting: the forecast package for R**. Journal of Statistical Software, 26(3):1–22, 2008.

R Core Team. R: **A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria, 2020.