Modelos ARIMA, SARIMA e Método de Seleção de Variáveis do tipo LASSO para Séries Temporais

Gladys Choque Ulloa[1]

[1]Universidade Federal do Rio Grande do Sul.
Insituto de Matemática e Estatística.

20 novembro 2022

Conteúdo

Conteúdo

Programa de Pós-Graduação EM ESTATÍSTICA IME-UFRGS

Introdução

- Uma série temporal é uma sucessão de dados medidos em determinados momentos e organizados cronologicamente.
- Os dados podem ser espaçados em intervalos iguais (como a temperatura em um observatório meteorológico em dias sucessivos ao meio-dia) ou desiguais (como o peso de uma pessoa em medições sucessivas no consultório médico, farmácia, etc.).
- Para a análise de séries temporais, são utilizados métodos que ajudam a interpretá-las e que permitem extrair informações representativas sobre as relações subjacentes entre os dados da série ou de diferentes séries e que permitem, em diferentes graus e com diferentes níveis de confiança, extrapolar ou interpolar os dados e assim prever o comportamento da série em tempos não observados, seja no futuro (extrapolação de previsão), no passado (extrapolação para trás) ou em tempos intermediários (interpolação).

Programa de Pós-Graduação EM ESTATÍSTICA IME-UFRGS

Introdução

- As séries temporais são estudadas em estatística, processamento de sinais, econometria e muitas outras áreas.
- Os modelos ARIMA e SARIMA são métodos usados para a análise de uma série temporal, onde cada um dos parâmetros dos modelos são usados para análise de dados.
- O método da seleção do melhor subconjunto de covariáveis se baseia em ajustar todos os modelos com k covariáveis e posteriormente escolher o melhor dentre eles com base em algum critério.

Conteúdo

Modelo ARMA (p, q)

- Os modelos Autorregressivos de Média Móvel, ou modelos ARMA (também chamados de modelos Box-Jenkins) são considerados modelos clássicos de séries temporais.
  Uma série temporal é um modelo ARMA(p, q) se satisfizer;

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j a_{t-j} + a_t \tag{1}$$

onde $\phi_1, \cdots, \phi_p$ são os parâmetros da parte AR, $\theta_1, \cdots, \theta_q$ são os parâmetros da parte MA e a_t é um ruído branco.
O modelo ARMA é uma ferramenta para prever valores futuros da série, e esta composto de duas partes, uma parte autoregressiva (AR) e uma parte de média móvel (MA).

Programa de Pós-Graduação EM ESTATÍSTICA IME-UFRGS

Modelo ARIMA (p, d, q)

- Uma série temporal Y_t segue um modelo de autoregressiva de médias móveis integrada, se a d-ésima diferença $W_t = \Delta^d Y_t$ é um processo ARMA estacionário. Se $\{W_t\}$ segue um modelo ARMA (p, q), dizemos que Y_t é um processo ARIMA(p, d, q). Para fins práticos, geralmente podemos tomar d $= 1$ ou no máximo 2. Suponha o seguinte modelo ARMA(p, q):

$$\Theta_p(L) Y_t = \Theta_q(L) a_t \tag{2}$$

onde o polinômio AR pode ser fatorado em termos de suas p raízes $L1, \cdots, Lp$

$$\Theta_p(L) = (1 - L_1^{-1} L)(1 - L_2^{-1} L) \cdots (1 - L_p^{-1} L) \tag{3}$$

Estratégia de Modelagem ARIMA

a) Identificação.
O primeiro passo desta fase é a análise gráfica dos dados com intuito de identificar e eliminar possíveis sazonalidades, presença de tendência determinística ou estocástica.

b) Estimação.
Após a escolha de um candidato, passamos para a estimativa dos parâmetros do modelo, isto é, $(\phi_1, \cdots, \phi_p \theta_1, \cdots, \theta_q)$.

c) Análise residual.
Nesta etapa utilizamos vários testes e gráficos para diagnosticar o ajuste do modelo proposto aos dados.

d) Predição.
Uma vez que concluimos que o modelo proposto se ajusta de forma adequada aos dados satisfazendo os pressupostos teóricos esperados, podemos obter previsões dentro e fora da amostra para o problema estudado.

Programa de Pós-Graduação EM ESTATÍSTICA IME-UFRGS

Modelo SARIMA

- O modelo SARIMA $(p, d, q) \times (P.D, Q)$ é definido como;

$$\Phi_p(B^s)\phi_p(B){\nabla_s}^D\nabla^d X_t = \Theta_Q(B^s)\Theta_q(B)\epsilon_t \tag{4}$$

Onde;

  - Generaliza todos os modelos da família ARIMA
  - Permite modelar séries estacionárias, bem como séries sazonais e não estacionárias,bem com séries sazonais e não sazonais.

Em um modelo ARIMA, os termos estão incluídos autoregressivos (p), diferenciação da variável (d) e termos de média móvel (d). No entanto, o modelo SARIMA inclui termos sazonais autorregressivos (P), diferenciação sazonal (D) e média móvel sazonal (Q), ou seja, o SARIMA contém fatores sazonais e não sazonais em um modelo multiplicativo (Dritsaki, 2016:137).

Programa de Pós-Graduação EM ESTATÍSTICA IME-UFRGS

- O estimador via LASSO no contexto da regressão linear é dado pela resolução do problema

$$\hat{\beta}_L = \operatorname{argmin}\left\{ \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \quad \lambda \geq 0, \tag{5}$$

em que λ é um valor que deve ser escolhido previamente. Note que, se $\lambda = 0$, então o estimador via LASSO será igual ao estimador via mínimos quadrados, isto é, o estimador de mínimos quadrados de β pode ser visto como um caso particular do estimador via LASSO.

Conteúdo

Programa de Pós-Graduação EM ESTATÍSTICA IME-UFRGS

Aplicação

- Para parte aplicativa foi usado o banco de dados do Instituto Nacional de Meteorologia do Ministério da Agricultura, Pecuária e Abastecimento, do estado do Rio Grande do Sul,Brasil, no período de 2018-2021. Primeiramente faremos um ajuste dos modelos ARIMA e SARIMA e depois aplicaremos um método de seleção de variáveis com o LASSO.

Ajuste do Modelo ARIMA

- Neste caso seguiremos os passos para modelagem de um Modelo ARIMA, esta metodologia é conhecida também como Box Jenkins. Vamos começar fazendo o análise exploratório de nossos dados, temos ás variáveis "Precipitação","Pressão","Temporvalho, "MaxTemp", "MeanTemp", "MinTemp", "UMIDADE", "UmidadeMin", "MaxVento"e "MeanVento".

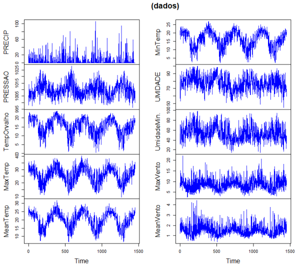

Figura: Séries Temporais das Variáveis de estudo.

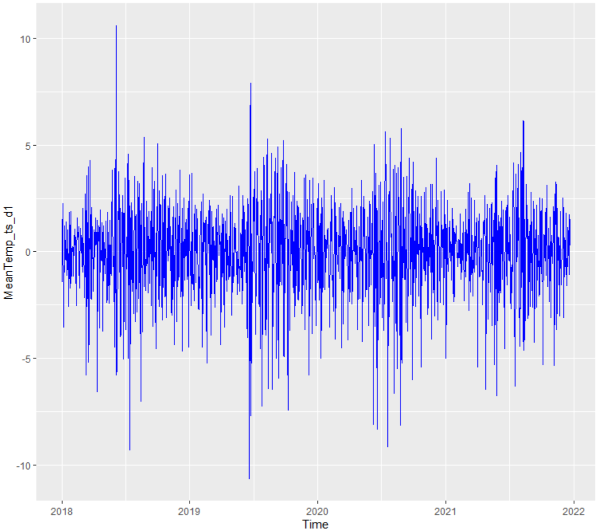

Figura: Séries Temporais das Variáveis de estudo.

- O gráfico apresenta o quão bem o modelo se ajusta aos dados. Nota-se por meio da figura que os resíduos apresentam uma distribuição aleatória sem apresentar tendência. Portanto, os resíduos são ruído branco, verificado através do teste de $Ljung - Box = 0.346$ e com $AIC = 6088.59$, na análise gráfica que os resíduos se distribuem de forma homogênea.

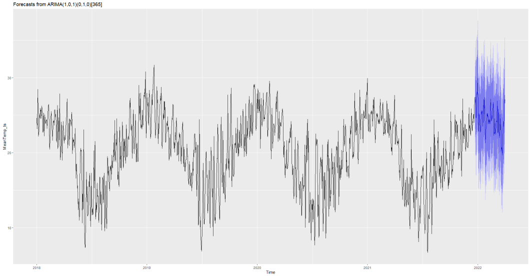

Figura: Gráfica de Previsão com o modelo Sarima.

Programa de Pós-Graduação EM ESTATÍSTICA
IME-UFRGS

Ajuste de uma Regressão com o LASSO

O processo para realizar um ajuste de Lasso e identificar o melhor valor lambda e o procedimento é equivalente ao caso Ridge.Para fazer esse ajuste usaremos a função glmnet() onde alfa=1.

Para a análise, primeiro precisamos de dois arquivos "Train" e "Teste1" do nosso banco de dados,onde suas dimensões devem ser iguais.
Na saída do R a quantidade dos Modelos gerados com o LASSO são 69 com 10 variáveis.

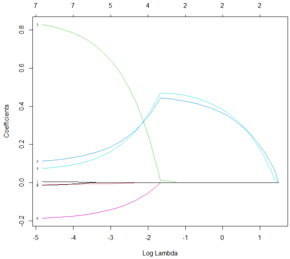

Figura: Gráfico dos valores de λ.

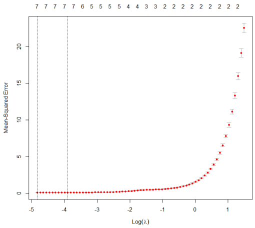

Figura: Gráfico do valor de λ.

Vemos no gráfico acima que o valor de melhor $\lambda = -4.836514$ está entre esse intervalo das linhas pontilhadas.
Após de fazer a previsão com o melhor λ e obter os coeficientes do modelo, podemos observar que as variáveis que mais contribuem para o modelo são Precip,Pressao,TempOrvalho,MaxTemp,MinTemp,Umidade,MaxVento e o Intercept.
Obtivemos um $RMSE = 0.2739275$, o que indica que é o melhor ajuste para nosso modelo.

Programa de Pós-Graduação EM ESTATÍSTICA IME-UFRGS

Conteúdo

Discussão

- Os modelos ARIMA e SARIMA são uma classe de modelos de Séries Temporais, que são geralmente usados para a análise de previsão.
- O modelo ARIMA é um método estatístico "clássico" para análise de séries temporais: detecção, previsão e previsão de anomalias. Para diversificar a sua utilização (por exemplo, trabalhar em tempo real) pode ser combinada com outras técnicas.
- Verificou-se que supera métodos mais complexos (ML) na previsão de dados univariados. Portanto, recomenda-se usá-lo como linha de base para demonstrar que a complexidade agregada por esses métodos agrega valor.

Conteúdo

Programa de Pós-Graduação EM ESTATÍSTICA IME-UFRGS

Referências

- Rodrigues, K. A. (2018). LASSO Clássico e Bayesiano. Instituto de Matemática e Estatística,Universidade de São Paulo, Brasil.
- LUCAS, J. L. (2013). Modelos de Series Temporais para Previsão da Demanda. Faculdade de Economía, Administração, Atúaria, Contabilidade e Secretariado Executivo, Brasil.
- Miranda, C. d. (2001). Modelação Linear de Séries Temporais na presença de . Departamento de Matemática Aplicada, Faculdade de Ciências da Universidade do Porto, Brasil.
- Konzen, E. (2014). Penalizações Tipo Lasso na Seleção de Covariáveis em Séries Temporais. Departamento de Economía, Universidade Federal do Rio Grande do Sul, Brasil.
- Freitas, A. A. (2007). Previsão de Séries Temporais via Seleção de Variáveis, Reconstrução Dinâmica, ARMA-GARCH e Redes Neurais Artificiais. Faculdade de Engenharia Elétrica e de Computação,Universidade Estadual de Campinas,Brasil.
- Valipour, M. (2015). Long-term runoff study using SARIMA and ARIMA models in the United States. Meteorological Applications, 22(3), 592-598.

Programa de Pós-Graduação EM ESTATÍSTICA IME-UFRGS

Referências

- Tseng, F. M., Yu, H. C., & Tzeng, G. H. (2002). Combining neural network model with seasonal time series ARIMA model. Technological forecasting and social change, 69(1), 71-87.
- Otu, O. A., Osuji, G. A., Opara, J., Mbachu, H. I., & Iheagwara, A. I. (2014). Application of Sarima models in modelling and forecasting Nigeria's inflation rates. American Journal of Applied Mathematics and Statistics, 2(1), 16-28.
- Kajuru, J. Y., Abdulkarim, K.,& Muhammed, M. M. (2019). Forecasting Performance of ARIMA and Sarima Models on Monthly Average Temperature of Zaria, Nigeria. ATBU Journal of Science, Technology and Education, 7(3), 205-212.
- Sun, K., Huang, S. H., Wong, D. S. H., & Jang, S. S. (2016). Design and application of a variable selection method for multilayer perceptron neural network with LASSO. IEEE transactions on neural networks and learning systems, 28(6), 1386-1396.
- Yan, Z., & Yao, Y. (2015). Variable selection method for fault isolation using least absolute shrinkage and selection operator (LASSO). Chemometrics and Intelligent Laboratory Systems, 146, 136-146.
- Ranstam, J., & Cook, J. A. (2018). LASSO regression. Journal of British Surgery, 105(10), 1348-1348.

Programa de Pós-Graduação EM ESTATÍSTICA IME-UFRGS