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

Gladys Choque Ulloa

Fevereiro 2022

Resumo

No presente trabalho se apresenta a teoria e a aplicação dos Modelos ARIMA, SARIMA e um método de seleção de variáveis como o LASSO, usando os testes de Dickey Fuller e Ljung Box, fazendo uso da linguagem de programação R para a estimativa dos coeficientes. Foi feito uma comparação desses dois modelos para determinar qual é o mais adequado para representar a série original e estimar os valores futuros, constatando que o modelo SARIMA tem melhor ajuste e melhor previsão de seu comportamento. Ridge e LASSO são métodos de penalização que são uma modificação das técnicas clássicas de regressão. A diferença entre Ridge e LASSO é que Ridge contrai os coeficientes, mas não os torna zero, em vez disso, LASSO torna os coeficientes zero quando não são tão importantes, esse método de penalidade também é conhecido como método de seleção de variáveis.

Palabras-Chave: ARIMA, SARIMA, LASSO, Forecasting.

1 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). Um dos usos mais comuns das séries temporais é sua análise para previsão (isso é feito, por exemplo, com dados meteorológicos, ações do mercado de ações ou séries de dados demográficos). É difícil imaginar uma área da ciência em que não apareçam dados que possam ser considerados como séries temporais. 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.

Existem casos em que o número de covariáveis é muito grande, ou mesmo, maior que o tamanho da amostra ($p > n$). Neste caso é necessário selecionar quais covariáveis serão utilizadas no modelo e quais não serão selecionadas. Os principais métodos para a seleção de variáveis são: seleção do melhor subconjunto de covariáveis, backward, forward, stepwise e LASSO (Least Absolute Shrinkage and Selection Operator). 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. Para aplicar este método de seleção de covariáveis é necessário definir o número de covariáveis do modelo, k e também definir o critério de comparação a ser utilizado a priori. Um dos métodos mais conhecidos é o LASSO, que é um método de seleção e redução de variáveis que possui a propriedade da regressão de cristas (ridge regression) de reduzir o valor das estimativas dos parâmetros mas é capaz de produzir estimativas iguais a zero para os parâmetros do modelo como no método da seleção do melhor subconjunto de covariáveis gerando assim modelos interpretáveis. Este método de seleção e redução de covariáveis foi adaptado para vários modelos como séries temporais (Audrino e Camponovo, 2013), processos autoregressivos com caudas pesadas (Sang e Sun, 2013) e também para regressão L_1 (Wang et al., 2007).

2 Metodología

2.1 Modelos Lineares Estacionários

2.1.1 Modelo Linear Geral

A metodologia de modelagem univariada é simple. Como o objetivo é explicar o valor tomado no tempo t por uma variável econômica que exibe dependência do tempo, uma maneira de trabalhar é reunir informações sobre o passado da variável, observar sua evolução ao longo do tempo e explorar o padrão de regularidade mostrado pelos dados. A estrutura da dependência do tempo de um processo estocástico é coletada na função de autocovariância (FACV) e/ou na função de autocorrelação (FAC).

Nesse contexto, trata-se de usar informações dessas funções para extrair um padrão sistemático e, a partir disso, um modelo que reproduz o comportamento da série e pode ser usado para prever. Este procedimento será operacionalizado por modelos ARMA que são uma aproximação a uma estrutura teórica geral. Em um modelo de série temporal univariada, a série Y_t é dividida em duas partes, uma que inclui o padrão de regularidade, ou parte sistemática, e outra parte puramente aleatória,também chamado de inovação.

2.1.2 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).

2.2 Modelos Lineares Não Estacionários

2.2.1 Não Estacionaridade na Variação

Quando uma série não é estacionária em variância, ou seja, sua variância varia ao longo do tempo, a solução é transformar a série por algum método que estabiliza a variância. O comportamento habitual em série econômica é geralmente que a variância muda à medida que o nível da série muda. Nestes casos, assumimos que a variância do processo pode ser expressa como alguma função do nível:

$$V(Y_t) = kf(\mu_t), \tag{2}$$

onde $k > 0$ é uma constante e f é uma função conhecida. O objetivo é conseguir alguns função que transforma a série para que $h(Y_t)$ tenha variância constante. Em geral, as transformações de Box-Cox são usadas para estabilizar a variância:

$$Y_t^{(\lambda)} = \begin{cases} \dfrac{Y_t^{\lambda} - 1}{\lambda}, & \text{se } \lambda \neq 0 \\ \ln(Y_t), & \text{se } \lambda = 0. \end{cases} \tag{3}$$

onde λ é o parâmetro de transformação. É interessante notar que as transformações Box-Cox geralmente não apenas estabilizam a variância, mas também melhoram a aproximação para a distribuição normal do processo Y_t.

2.2.2 Não Estacionaridade na Média

Uma das características dominantes e facilmente observáveis em séries temporais é a presença de tendência. A tendência é o movimento de longo prazo da série uma vez que os ciclos e o prazo irregular são eliminados. Dentro da economia esta tendência é geralmente produzida devido à evolução das preferências, tecnologia, demografia, etc. Esse comportamento de tendência pode ser crescente ou decrescente, exponencial ou aproximadamente linear. As séries que apresentam um comportamento sistemático deste tipo não são estacionários, eles não evoluem em em torno de um nível constante.

A não estacionaridade na média pode ser modelada de diferentes maneiras. Por um lado, é possível modelar tendências do modelo, mudanças sistemáticas de nível, por meio de modelos globais em que especifica a tendência em função do tempo:

$$Y_t = T_t + \mu_t \tag{4}$$

onde $T_t = f(t)$ é uma função determinística do tempo e μ_t é um processo estocástico estacionário com média zero.

2.2.3 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):

$$\Theta_p(L)Y_t = \Theta_q(L)a_t \tag{5}$$

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

$$\Theta_p(L) = (1 - L_1^{-1}L)(1 - L_2^{-1}L) \cdots (1 - L_p^{-1}L) \tag{6}$$

Suponha que $(p - 1)$ raízes sejam estacionárias (com módulo fora do círculo unitário) e um deles é unitário, $L_i = 1$. Então, o polinômio AR pode ser reescrito da seguinte forma:

$$\begin{aligned} \Theta_p(L) &= (1 - L_1^{-1}L)(1 - L_2^{-1}L) \ldots (1 - L_p^{-1}L) \\ &= \varphi_{p-1}(L)(1 - 1_1^{-1}L)\Theta_p(L) \\ &= \varphi_{p-1}(L)(1 - L) \end{aligned} \tag{7}$$

onde o polinômio $\varphi_{p-1}(L)$ resulta do produto dos $(p - 1)$ polinômios de ordem 1 associados às raízes L_i com módulo fora do círculo unitário. Substituindo no modelo ARMA(p, q) na equação(6) temos:

$$\varphi_{p-1}(L) = (1 - L)Y_t = \Theta_q(L)a_t\varphi_{p-1}(L)\Delta(Y_t) = \Theta_q(L)a_t \tag{8}$$

onde o polinômio $\varphi_{p-1}(L)$ é estacionário porque suas raízes têm módulo fora do círculo unitário e o polinômio $\Delta = (1 - L)$, de orden d, contém as raízes unitárias não estacionárias.

O modelo (8) representa o comportamento de um processo Y_t que não é estacionário porque tem raiz unitária. Um processo Y_t com essas características é chamado de processo integrado de ordem 1.

Definição 2.1. Se um processo Y_t é um processo $I(d)$ tal que $\Delta^d Y_t$ é um processo ARMA estacionário, então pode ser representado como

$$\phi_p(L)\Delta^d Y_t = \Theta_q(L)a_t. \tag{9}$$

onde a_t é um ruído branco e assumimos ainda que o polinômio autoregressivo $\phi_p(L)$ não possui raízes unitárias e os polinômios AR e MA não possuem raízes em comum. Um processo satisfazendo (9) é chamado de ARIMA(p, d, q)

2.2.4 Estratégia de Modelagem ARIMA

A construção dos modelos ARIMA é realizada de forma iterativa através de um processo em quais quatro fases podem ser distinguidas:

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,

entre outras características nos dados que precisam ser tratadas antes da modelagem ARMA, seja via a diferenciação da série, regressão polinomial (para eliminar tendências determinísticas) ou em variáveis dummy (para eliminar picos que se repetem), etc. Após esta etapa, o objetivo é tentar identificar um candidato inicial dentro da classe de modelos ARMA(p, q), isto é, identificar p e q, para os dados. Esta etapa é realizada com o auxílio dos gráficos da ACF e PACF dos dados. Pode ocorrer de mais de um modelo inicial ser identificado para os dados.

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,verificando se os pressupostos do modelo estão satisfeitos e propondo novos modelos casos a aderência do modelo atual aos dados seja pobre.

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.

2.3 Modelo SARIMA

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

$$\Phi_p(B^s)\phi_p(B)\nabla_s{}^D\nabla^d X_t = \Theta_Q(B^s)\Theta_q(B)\epsilon_t \tag{10}$$

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).

2.4 Modelo Linear

O modelo de regressão linear é dado por

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{pi} + \epsilon_i, \quad i = 1, 2, \cdots, n. \tag{11}$$

em que Y_i é a i-ésima variável resposta, β_0 é o intercepto do modelo, β_j, $j = 1, \cdots, p$, é o parâmetro associado à j-ésima covariável, X_{ji}, $j = 1, \cdots, p$, $i = 1, \cdots, n$ é a j-ésima covariável da i-ésima observação e ϵ_i é o erro associado à i-ésima observação. Podemos expressar o modelo de forma mais compacta, como:

$$Y = X\beta + \epsilon, \tag{12}$$

onde, $Y_{n\times 1} = (Y_1, \cdots, Y_n)^T, X_{n\times(p+1)} = ({X_1}^T, \cdots, {X_p}^T)^T$, $X_k = (1, X_{1k}, \cdots, X_{pk})$, $\beta_{(p+1)\times 1} = (\beta_0, \cdots, \beta_p)^T$ e $\epsilon_{n\times 1} = (\epsilon_1, \cdots, \epsilon_n)^T$, em que $\epsilon \sim N(0, I_n\sigma^2)$.

2.5 Conceitos Básicos Sobre o LASSO

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

$$\hat{\beta}_L = \text{argmin}\left\{ \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \quad \lambda \geq 0, \tag{13}$$

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.

O estimador via LASSO é basicamente o problema de minimização usual dos erros ao quadrado acrescido de uma penalização L_1 em relação aos parâmetros de posição do modelo de regressão. Um detalhe importante acerca da penalidade é que o parâmetro β_0 não é restringido pela penalização. Note que quanto maior for o valor de λ maior será a penalização, ou seja, quanto maior for λ mais o vetor $\hat{\beta}_L$ se aproximará do vetor $(\hat{\beta}_0, 0^T)^T$, pois, se $\lambda \to \infty$, a penalização tenderá ao infinito, isto é,

$$\lim_{\lambda\to\infty} \lambda \sum_{j=1}^{p} |\beta_j| = \infty \tag{14}$$

e, consequentemente, todas as estimativas dos parâmetros associados às covariáveis serão excluídas do modelo; assim teremos um modelo somente com intercepto.

De forma geral, este problema não possui solução analítica como o problema de mínimos quadrados. Portanto, é necessário utilizar algoritmos para obter a estimativa via LASSO em problemas reais. Um caso bem conhecido em que o estimador assume forma fechada é o caso em que a matriz de especificação do modelo, X, é ortonormal.

Outra forma de escrever o problema de minimização dado em (13) é

$$\hat{\beta}_L = \text{argmin}\left\{ \sum_{i=1}^{n} (Y_i - X_i\beta)^2 \right\}, \quad \text{restrito a} \quad \sum_{j=1}^{p} |\beta_j| \leqslant s, \ s \geq 0. \tag{15}$$

Note que λ não está presente em (15) e surgiu um s. A relação entre λ e s é de grandezas inversamente proporcionais, ou seja, quando o valor de λ aumenta o valor de s diminui e vice-versa. De forma mais clara, considere duas aplicações do LASSO, uma com $\lambda = \lambda_1$, cujo problema pode ser escrito na forma de (19) com $s = s_1$, e outra com $\lambda = \lambda_2$ e $s = s_2$. Então,vale

$$\lambda_1 > \lambda_2 \Longleftrightarrow s_1 < s_2$$

Uma observação importante é que, em (19), se $s \geqslant \sum_{j=1}^{p} |\hat{\beta}_j|$, onde os $\hat{\beta}_j$'s são as estimativas de mínimos quadrados dos parâmetros , então o estimador via LASSO de β coincidirá com o estimador de mínimos quadrados de β, porque, neste caso, não há

restrição para as estimativas dos parâmetros. Por outro lado, se $s = 0$, então somente a estimativa do intercepto será não nula.

Antes de discutirmos uma das mais importantes características do estimador LASSO apresentaremos o estimador via ridge regression que é dado por

$$\hat{\beta_R} = \text{argmin}\left\{ \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \eta \sum_{j=1}^{p} {\beta_j}^2 \right\}, \ \eta \geq 0. \tag{16}$$

De modo análogo ao estimador via LASSO, podemos reescrever (16) na forma

$$\hat{\beta_R} = \text{argmin}\left\{ \sum_{i=1}^{n} (Y_i - X_i\beta)^2 \right\}, \quad \text{restrito a} \quad \sum_{j=1}^{p} {\beta_j}^2 \leqslant r, \ r \geq 0 \tag{17}$$

Considerações similares às apresentadas para o parâmetro λ do LASSO podem ser feitas sobre o parâmetro η da ridge regression. O estimador via ridge regression foi proposto antes do LASSO surgir. Perceba que a única diferença entre as equações (13) e (16) é a forma da penalização; enquanto a penalização L_1 é utilizada no LASSO, a penalização L_2 é usada na ridge regression. Essa pequena mudança interfere na capacidade das estimativas dos parâmetros do modelo efetivamente poderem ser nulas.

A Figura 1 apresenta a representação gráfica das equações (13) (no lado esquerdo da figura) e (16) (no lado direito da figura) no caso em que temos um total de duas covariáveis.

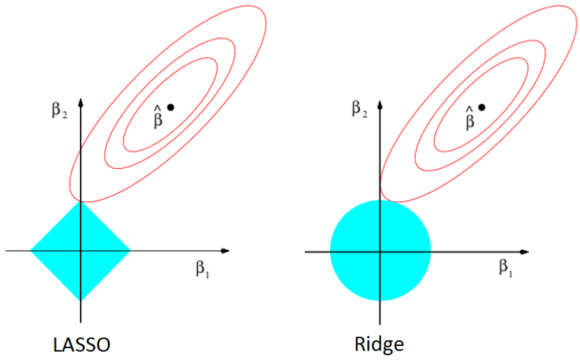

Figura 1: : Esquema gráfico do LASSO e Ridge regression no caso em que existem apenas duas variáveis explicativas.

Fonte: Hastie, Tibshirani e Friedman (2008) (com adaptações).

2.6 Como Escolher λ

Escolher um valor de λ é fundamental para que o método funcione adequadamente. Lembre que, se $\lambda = 0$, então o método de estimação é simplesmente o método de mínimos quadrados e se escolhemos um valor extremamente grande para λ então provavelmente acabaremos com um modelo sem covariáveis, isto é, só possui intercepto. Um modo de escolher o valor de λ é a validação cruzada, que é um método bastante fácil de entender e de implementar. Basicamente dividiremos, aleatoriamente, a amostra em k partes iguais ou pelo menos aproximadamente iguais.

Escolheremos a primeira parte para ser os dados de "validação" e as demais para ser dados de "treinamento", ajustaremos o LASSO com $\lambda = \lambda_0$ aos dados de "treinamento" e usaremos esse modelo para tentar prever os dados de "validação", então calcularemos o erro de predição. O mesmo procedimento será repetido mais $k-1$ vezes para as outras partes restantes. Após terminar as k iterações teremos k erros de predição e calcularemos a média dos k erros de predição. Esse procedimento será feito para vários valores λ_0 distintos. Finalmente, escolheremos o valor λ_0 que minimize o erro de predição médio.
É importante salientar que este método de escolha do valor de λ não depende do conhecimento do número de parâmetros ("graus de liberdade") do modelo e também não depende de uma estimativa do parâmetro de escala do modelo. Embora na situação tradicional em que se usa o estimador de mínimos quadrados, o número de parâmetros do modelo ajustado esteja bem definido e seja fácil achar uma estimativa para o parâmetro de escala, no caso do LASSO isso não é trivial, pois λ também é uma quantidade que tem impacto no ajuste do modelo.
O erro de predição relacionado a λ_0 é dado por

$$EP_{\lambda_0} = \frac{1}{n}\sum_{i=1}^{k} n_i EQM_i, \tag{18}$$

em que n_i é o número de observações da i-ésima parte dos dados e

$$EQM_i = \frac{1}{n_i}\sum_{j=1}^{n_i} (Y_j - \hat{Y}_j)^2, \tag{19}$$

ou seja, EQM_i é o erro quadrático médio quando usamos a i-ésima parte dos dados como "validação" e o erro de predição é simplesmente uma média ponderada dos EQM's. Essa é a quantidade que nos guiará na escolha do valor de λ; quanto menor o erro de predição, mais adequado é o valor de λ.

2.7 Inferência no LASSO

A inferência no LASSO ainda é um tópico de pesquisa aberto. O tópico de pesquisa que engloba esse tipo de problema se chama "inferência após seleção de variáveis". A razão para o surgimento desta área é que ao selecionarmos as covariáveis do modelo estamos influenciando a signifiância dos testes de hipóteses e o valor-P.

2.8 Banco de Dados

Para este trabalho usaremos 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.

3 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.

3.1 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".

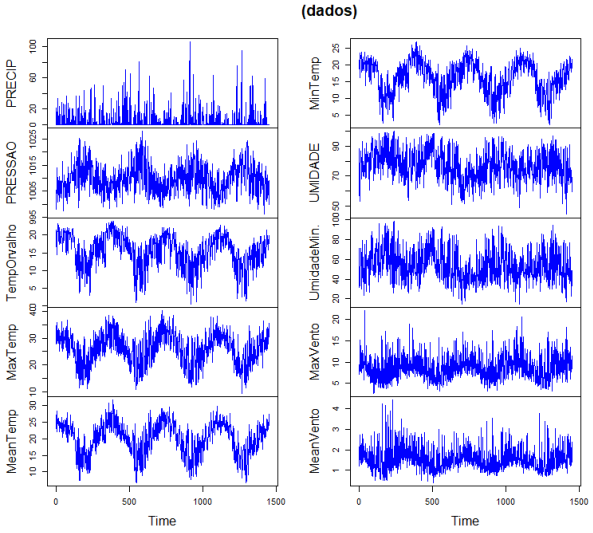

Figura 2: Séries Temporais das Variáveis de estudo.

Vemos do gráfico acima ás gráficas das séries temporais com relação à variável tempo e a cada uma das variáveis de nosso estudo.

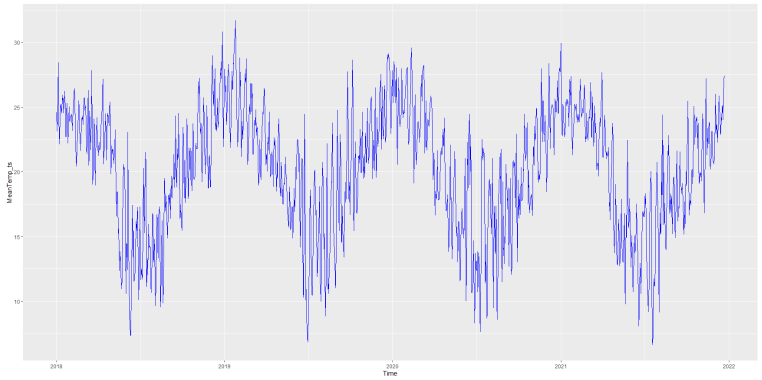

Figura 3: Série Temporal de nosso Banco de Dados.

Uma vez plotada a série temporal com ás variavéís MeanTempts e Tempo, percebe-se que ela apresenta comportamento sazonalidade. A partir da função auto.arima e

considerando o penalizador de Critério de Informação de Akaike (AIC) foi identificado que o modelo ARIMA(1,0,1)(0,1,0)[365] apresenta o melhor ajuste a série temporal de temperatura média.

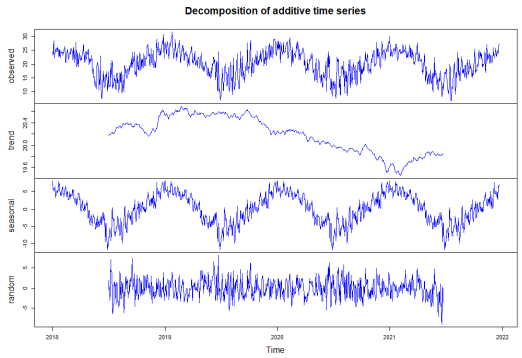

Figura 4: Série Temporal e sua decomposição.

Da figura 4, podemos ver a decomposição da série temporal em 4 componentes (Observed,trend,seasonal e random), onde também podemos observar o comportamento sazonal da nossa série temporal.

Agora vamos fazer os gráficos ACF e PACF para ver quantas médias móveis e quantos autorregressivos vamos usar em nosso modelo ARIMA.

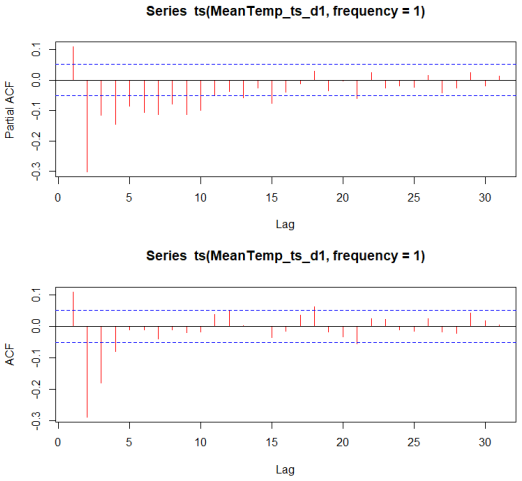

Figura 5: Função de auto-correlação e auto-correlação parcial.

A figura 5 se refere às funções de auto-correlação e autocorrelação parcial da temperatura média no estado do Rio Grande do Sul. Essa figura indica que a série possui certa dependência, que pode ser notada principalmente no gráfico da função de autocorrelação, além disso a série é não estacionaria. Os Lags para nosso banco de dados,em qualquer uma dessas gráficas, a faixa entre as linhas azuis intermitentes indica a zona de valores de correlação não significativas, que para fins práticos equivale a zero ou nenhuma correlação. Ao contrário, na área acima ou abaixo dessas bandas, a correlação é significativa, também as defasagens coincidem com as frequências.

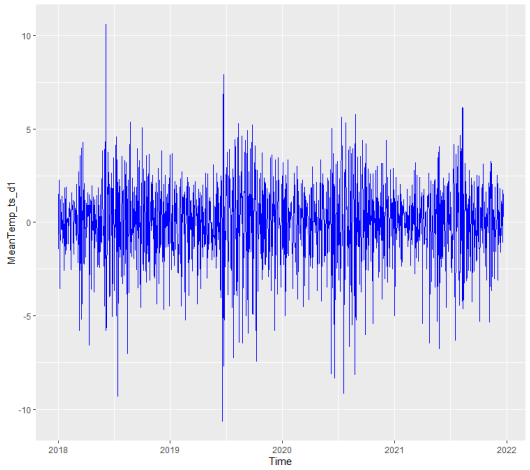

Figura 6: Série Temporal com diferença.

Da figura 6 podemos ver que os dados estão em torno da média zero e que a série é estacionária, mas para ter certeza disso, usaremos o teste de Dickey Fuller.

```
        Augmented Dickey-Fuller Test

data:  MeanTemp_ts_d1
Dickey-Fuller = -4.8422, Lag order = 365, p-value = 0.01
alternative hypothesis: stationary
```

Figura 7: Teste de Dickey Fuller.

A partir do resultado acima, podemos ver que nosso valor de p-value é $0,01 < 0.05$, o que nos diz que a série temporal é estacionária.

A partir da função auto.arima considerando o penalizador de Critério de Informação de Akaike (AIC) foi identificado que o modelo ARIMA(1,0,1)(0,1,0)[365] apresenta o melhor ajuste a série temporal da temperatura média.

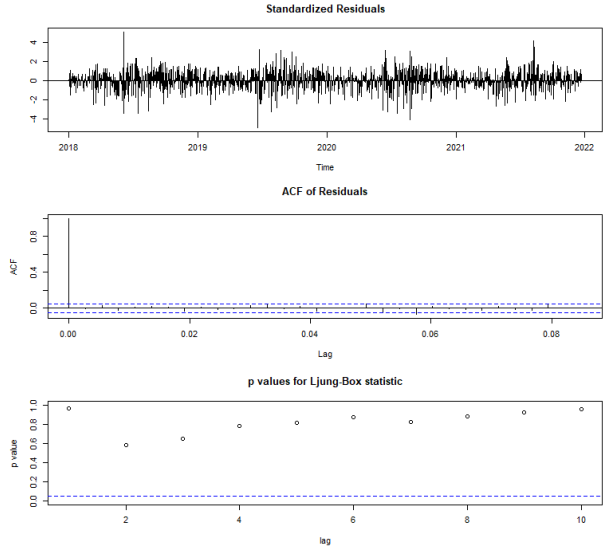

Figura 8: Pressuposições para o Modelo Sarima.

Da figura 8, além da ACF dos resíduos padronizados é possível verificar a homogeneidade dos mesmos.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.

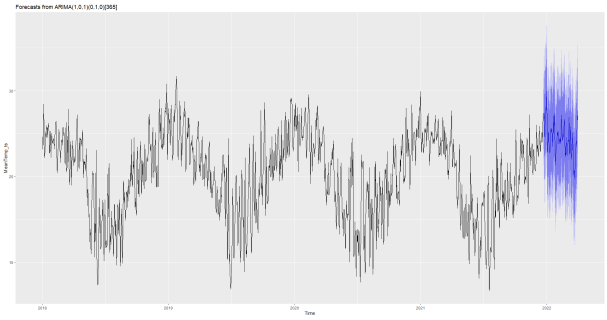

Figura 9: Gráfica de Previsão com o modelo Sarima.

Do gráfico acima vemos que se ilustra o ajuste de tal modelo aos dados da temperatura média do estado do Rio Grande do Sul. Nota-se que os valores ajustados estão bem próximos dos reais, o que significa que o modelo conseguiu captar a dinâmica temporal da série.

3.2 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.

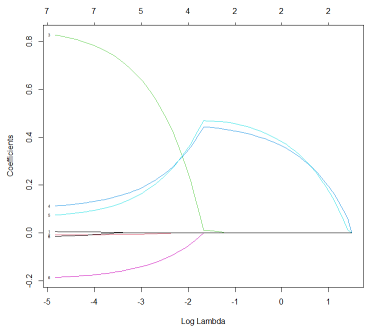

Figura 10: Gráfico dos valores de λ.

Vemos no gráfico acima os valores de λ gerados com o LASSO para os 69 modelos.

Agora temos que selecionar qual é o melhor λ e o melhor modelo para fazer nossa previsão, então vamos a fazer uma previsão com o LASSO para o modelo 60.

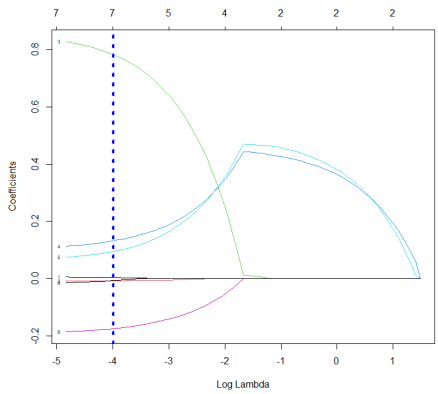

Figura 11: Valor do melhor λ para o Modelo 60

Vemos no gráfico acima que o valor de λ para o modelo 60 é 0.01833011 e aplicando o logaritmo é -3.99921, também temos os valores para os coeficientes do mesmo modelo, além disso podemos observar que as variáveis que mais contribuem para o modelo são Precip,Pressao,TempOrvalho, MaxTemp,MinTemp,Umidade, MaxVento e o Intercept.

Agora vamos a fazer uma previsão para o modelo 15.

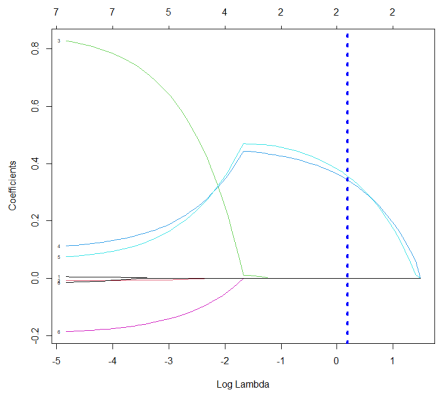

Figura 12: Valor do melhor λ para o Modelo 15.

Vemos no gráfico acima que o valor de λ para o modelo 15 é 1.205999 e aplicando o logaritmo é 0.1873084, também temos os valores para os coeficientes do mesmo modelo,além disso podemos observar que as variáveis que mais contribuem para o modelo são MaxTemp,MinTemp e o Intercept.

Agora vamos selecionar o melhor λ com a validação cruzada. Na saída do R o valor do melhor λ é 0.007934669 e aplicando o logaritmo é -4.836514.

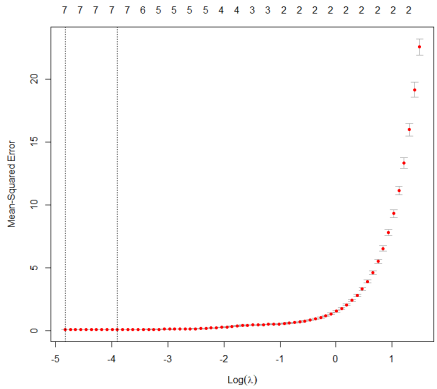

Figura 13: Gráfico do valor de λ.

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.

4 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.

A regressão de LASSO é o que chamamos de método de regressão penalizado, usado frequentemente em aprendizagem de máquina para selecionar um subconjunto de variáveis. É um método supervisionado de aprendizagem de máquina. Especificamente, LASSO é um método de Shrinkage e Seleção de Variáveis para modelos de regressão linear. LASSO, é na verdade um acrônimo para "Least Absolute Selection and Shrinkage Operator".

Para nosso estudo, aplicamos o modelo ARIMA,SARIMA e o método de seleção de varáveis com o LASSO para nosso banco de dados, onde foi obtido resultados perfeitos para cada modelo feito.

5 Referências Bibliográficas

5.1 Bibliografia

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.

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.