Curso de Vigilância em Saúde apoiado por plataforma BI

Aula 07- Analisando séries temporais com regressão binomial negativa

1. Apresentação

Olá!

Seja bem-vindo à quarta aula do curso de Vigilância em Saúde apoiado por plataforma *Business Intelligence* (BI)! Nesse curso estamos aprendendo alguns conceitos de Vigilância Epidemiológica apoiada por ferramentas de BI para análise de dados temporais.

O objetivo do nosso curso é ajudar a análise epidemiológica em um contexto de vigilância em serviços de saúde, utilizando ferramentas que facilitem a vida dos profissionais de saúde.

Essa é a aula sobre análise de séries temporais. Nessa aula, vamos usar dados do Sistema de Informações Hospitalares- SIH para ajustarmos um modelo binomial negativo. Essa é a primeira parte da aula e a segunda parte ocorrerá na próxima aula.

2. Introdução

Séries Temporais podem ser definidas como um conjunto de observações ordenadas no tempo. Isso significa que quando temos uma variável ocorrendo de forma periódica no tempo, temos uma série temporal (1). O conceito de série temporal é muito importante no campo da Estatística e pode ser amplamente aplicado na Epidemiologia e na Vigilância em Saúde.

Com uma infinidade de sistemas de informação, temos cada vez maior facilidade para a coleta de dados temporais. Com o monitoramento ininterrupto e a coleta de dados se ampliando, aumentará a necessidade e a possibilidade de análise de séries temporais (2).

Quando falamos de análises de séries temporais, estamos nos referindo a modelos estatísticos que nos permitem mostrar o comportamento dos dados e predizer quais serão seus valores no futuro. Isso pode ser realizado por uma infinidade de análises. Nesse curso vamos apresentar o modelo de regressão binomial negativo.

3. Modelo Binomial Negativo

O modelo de regressão binomial negativo é enquadrado nos modelos para dados de contagem, ou seja, é aplicado aos dados com valores inteiros e não negativos (3). Nesse caso ele será muito útil para a análise de incidência de casos novos no tempo.

Quando os dados são contados em um determinado intervalo de tempo, podemos modelá-los com base em uma distribuição binomial negativa, também conhecida como Poisson-Gama. Isso vai nos ajudar a compreender qual o comportamento dos dados além de predizer os valores que devem ocorrer no futuro da série (3).

Vamos ajustar um modelo binomial negativo com os dados de internação do SIH e utilizaremos esse exemplo para realizar a avaliação do modelo com uma programação no R. Na próxima aula, vamos utilizar o modelo ajustado para prever a incidência de novas internações e para auxiliar na identificação de um surto.

4. Analisando séries temporais com R - Parte 1

Você deve ter aberto o R pelo arquivo Rproject que está disponível em nossa plataforma. Caso não se lembre de como realizar essa operação, retorne à aula 4 com as instruções detalhadas.

Vamos iniciar carregando pacotes. Primeiro, vamos instalar o pacote *Pacman*, que permite a instalação e carregamento de quaisquer pacotes que estejam no CRAN.

Nós vamos ajustar um modelo binomial negativo, para isso, precisaremos de uma série de pacotes que trabalham com séries temporais.

```
#instalando e carregando pacotes
install.packages("pacman")

#carregando pacotes necessários
pacman::p_load(
   read.dbc,       # importando dados
   here,           # alocando arquivos
   tidyverse,    # manipulação de dados e criação de gráficos
   tsibble,        # manipulando séries temporais
   slider,         # calculo de médias móveis
   imputeTS,       # para corrigir dados perdidos
   feasts,         # para composição e autocorrelação de uma
série temporal
   forecast,       # ajustando predições temporais
   trending,       # ajuste e avaliação de modelos
   yardstick,      # para avaliar a acurácia do modelo
   surveillance    # para detectar dados errados
)
```

Após a importação dos pacotes, vamos importar os dados de internações do Distrito Federal. Vamos usar o número absoluto de internações por dia, pois no modelo binomial negativo trabalhamos com dados de contagem inteiros e não negativos.

```
#importando dados de internação

files            =            cbind(paste0("dados/sih_df/",
list.files("dados/sih_df/", pattern = "\\.dbc$")))

sihdf = NULL

for (i in 1:length(files)) {
```

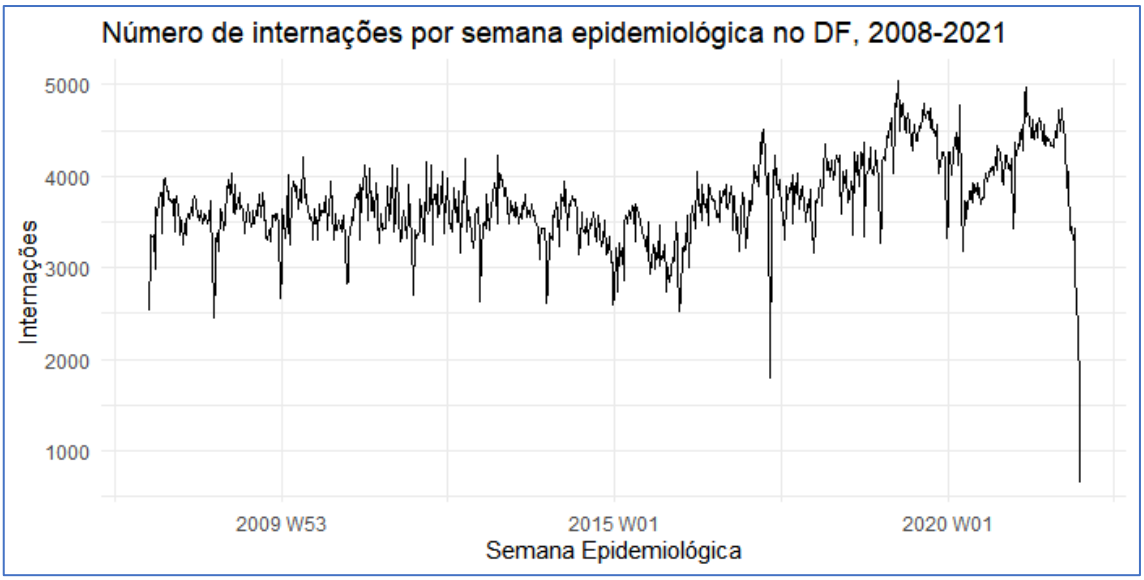
```
  sihdf_temp = read.dbc(files[i])
  sihdf_temp2          =          data.frame(DATA          =
as.Date(sihdf_temp$DT_INTER, "%Y%m%d"))
  sihdf= data.frame(rbind(sihdf, sihdf_temp2))
}
```

Vamos excluir os dados de *data.frames* que não usaremos, criaremos uma variável de contagem e vamos agregar o número de casos por dia. Depois disso, vamos usar a função *summary* para visualizar as medidas descritivas dos dados e vamos filtrar os dados para os anos entre 2008 e 2021.

Nós vamos agrupar as contagens por semana epidemiológica, para criarmos uma série com base na contagem de internações por semana.

```
#excluindo data.frames que não serão usados
sihdf_temp = NULL
sihdf_temp2 = NULL

#criando data.frame com internações por data
sihdf$CASOS = 1

#agregando número de casos por dia
sihdf2 = aggregate(CASOS ~ DATA, data = sihdf, sum)

#investigando a amplitude das datas
summary(sihdf2$DATA)

#definindo intervalo de tempo
sihdf2  =  subset(sihdf2,  sihdf2$DATA  >=  "2008-01-01"  &
sihdf2$DATA <= "2021-12-31")

#criando uma variável de semana epidemiológica
sihdf2 <- sihdf2 %>%
  mutate(SE = yearweek(DATA, week_start = 1))

sihdf3 = aggregate(CASOS ~ SE, data = sihdf2, sum)

#definindo  uma  série  temporal  a  partir  da  semana
epidemiológica
tssihdf <- tsibble(sihdf3, index = SE)
```

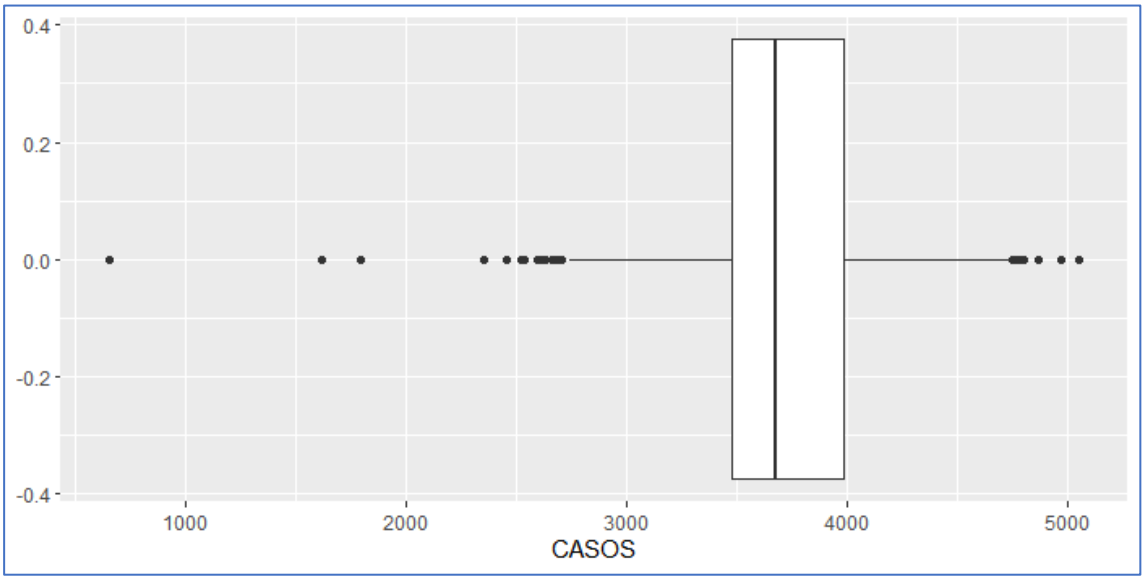
Após isso, nós vamos plotar a série para visualizar as internações por semana.



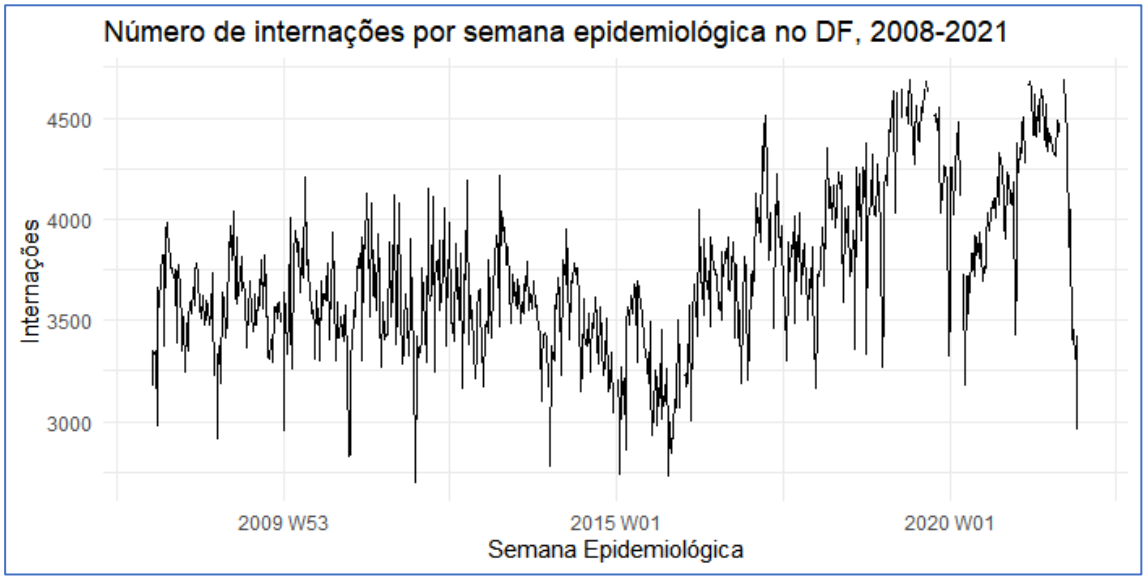
Nós percebemos que há pontos muito mais baixos que o restante da série. Por isso, vamos remover os *outliers* para não enviesar nossas análises. Vamos usar um gráfico *boxplot* para realizar a identificação visual dos *outliers*.

```
#removendo outliers

##identificando o outlier
ggplot(tssihdf, aes(x = CASOS))+
  geom_boxplot()

summary(tssihdf$CASOS)

tssihdf$CASOS = ifelse(tssihdf$CASOS < 2700 | tssihdf$CASOS
> 4700, NA, tssihdf$CASOS)
```

Recebemos a seguinte visualização. Perceba que os pontos considerados *outliers* são aproximadamente menores que 2.700 e maiores que 4.700. Por isso, vamos substituir esses dados por NA e plotar a série novamente.



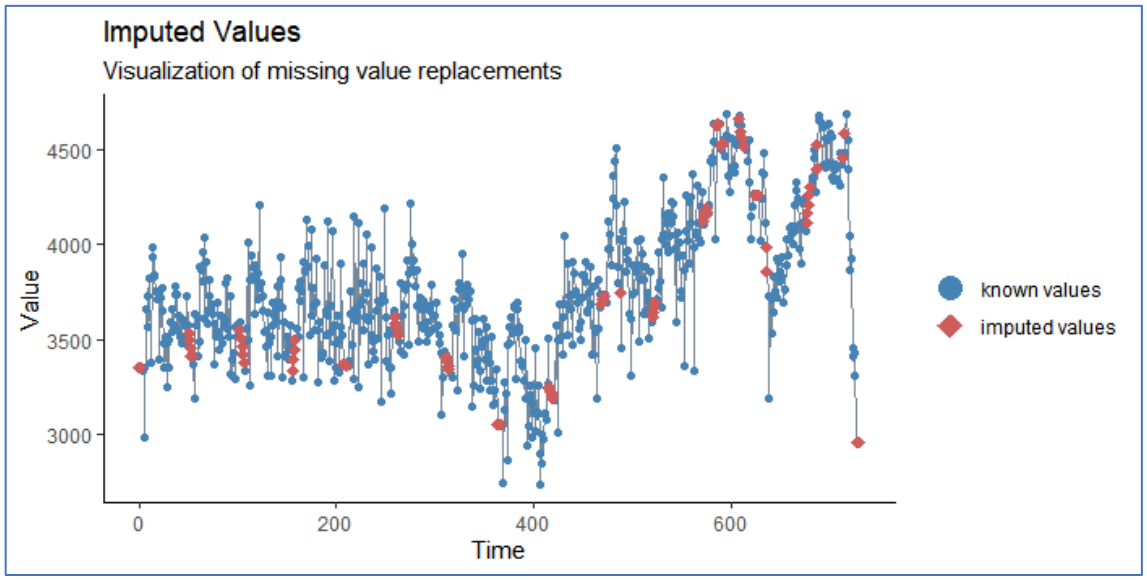
Após isso, vamos plotar a série novamente sem os *outliers*. Perceba que os dados estão bem mais aproximados entre si, sem grandes alterações.



Nós vamos realizar a interpolação dos dados para as semanas finais do ano, onde geralmente os dados estão mais defasados. Esse exercício é apenas para demonstrar a técnica de interpolação, pois os dados não aparentam lacunas nas semanas finais e iniciais do ano. Assim, faremos uma interpolação dos dados, substituindo os dados que estão vazios e depois substituindo-os por interpolação. Após isso, plotaremos o gráfico.

```
#identificando valores perdidos na série
tssihdf <- tssihdf %>%
  mutate(MISSING = if_else(
    str_detect(SE, "W51|W52|W53|W01|W02"),
    NA_real_,
    CASOS
  ))

#substituindo valores perdidos por interpolação
tssihdf <- tssihdf %>%
  mutate(CASOS_INT = imputeTS::na_interpolation(MISSING)
  )

##gerando gráfico com os dados interpolados
ggplot_na_imputations(tssihdf$MISSING, tssihdf$CASOS_INT) +
  theme_classic()
```

Para o gráfico, receberemos o seguinte *output*. Os pontos em vermelho foram interpolados para que possamos trabalhar com uma série mais próxima da realidade.

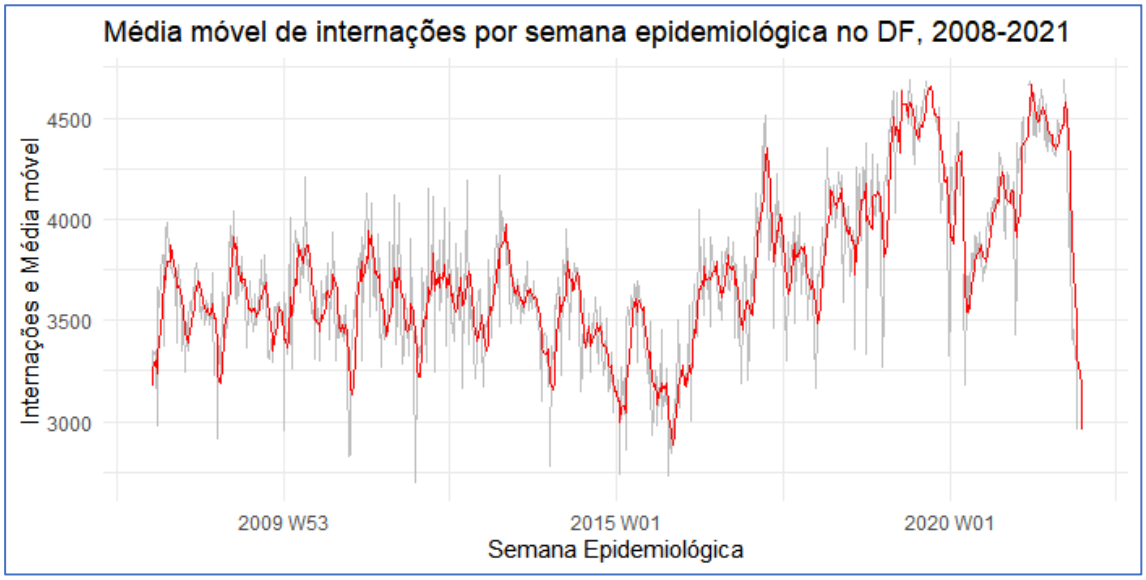


Nós vamos realizar a média móvel diretamente no gráfico, como mostramos na aula anterior. Essa técnica de suavização é boa para compreendermos o comportamento da série.

```
#analise descritiva da série temporal por médias móveis
tssihdf <- tssihdf %>%
  mutate(MA4SE = slider::slide_dbl(CASOS,
                                   ~ mean(.x, na.rm = TRUE),
                                   .before = 4))

##gerando um gráfico com a série e com a média móvel
```

```
ggplot(tssihdf, aes(x = SE)) +
  geom_line(aes(y = CASOS), colour = "grey") +
  geom_line(aes(y = MA4SE), colour = "red")+
  theme_minimal()+
  labs(
    title = "Média móvel de internações por semana epidemiológica no DF, 2008-2021",
    x = "Semana Epidemiológica",
    y = "Internações e Média móvel")
```

Recebemos o seguinte *output* para a série entre 2008 e 2021.

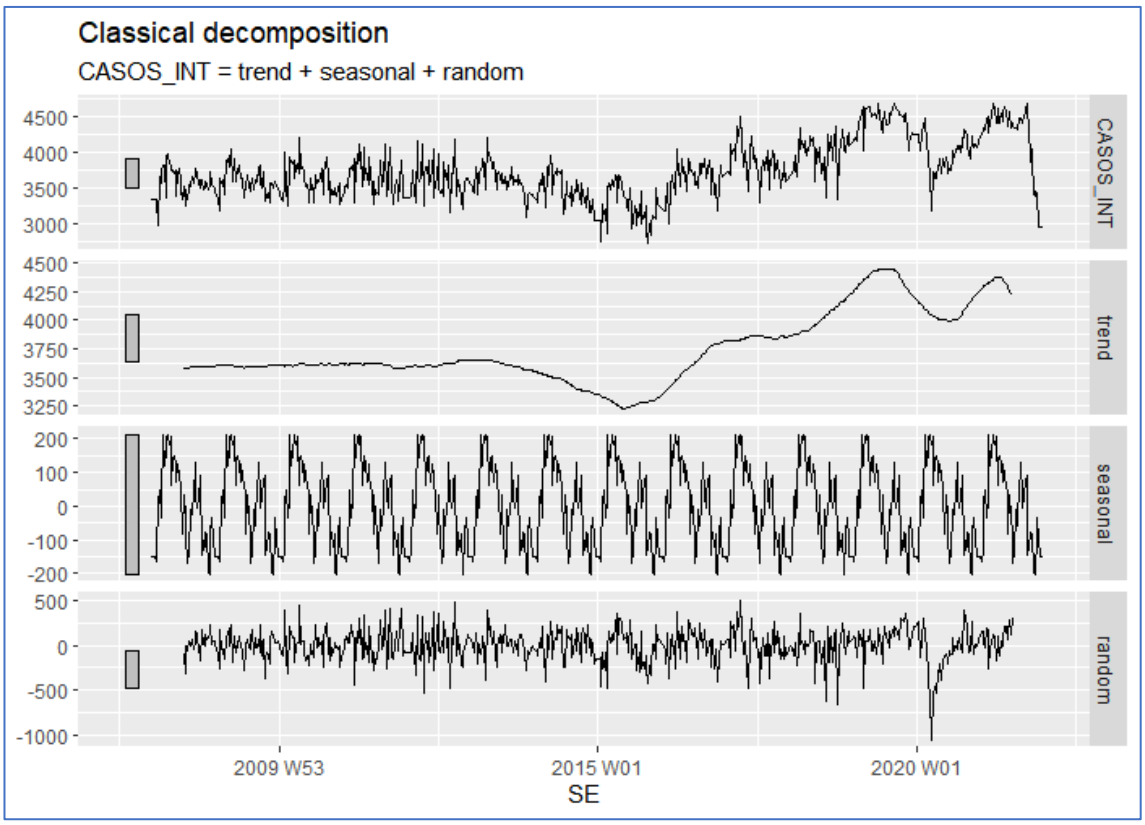


Agora seguiremos para a compreensão da série temporal. Inicialmente faremos a decomposição da série. Uma série temporal pode ser aditiva ou multiplicativa e pode possuir componentes de sazonalidade, tendencia, ciclicidade e um componente de erro aleatório.

```
#analisando os componentes da série temporal para análise inferencial

##decompondo a série temporal
tssihdf %>%
  model(classical_decomposition(CASOS_INT, type = "additive")) %>%
  components() %>%
  autoplot()
```

No caso da série que estamos analisando, há um componente de tendencia, que se mostra linear no início da série, apresenta uma queda e depois uma oscilação com aumento no final da séria. Há um forte

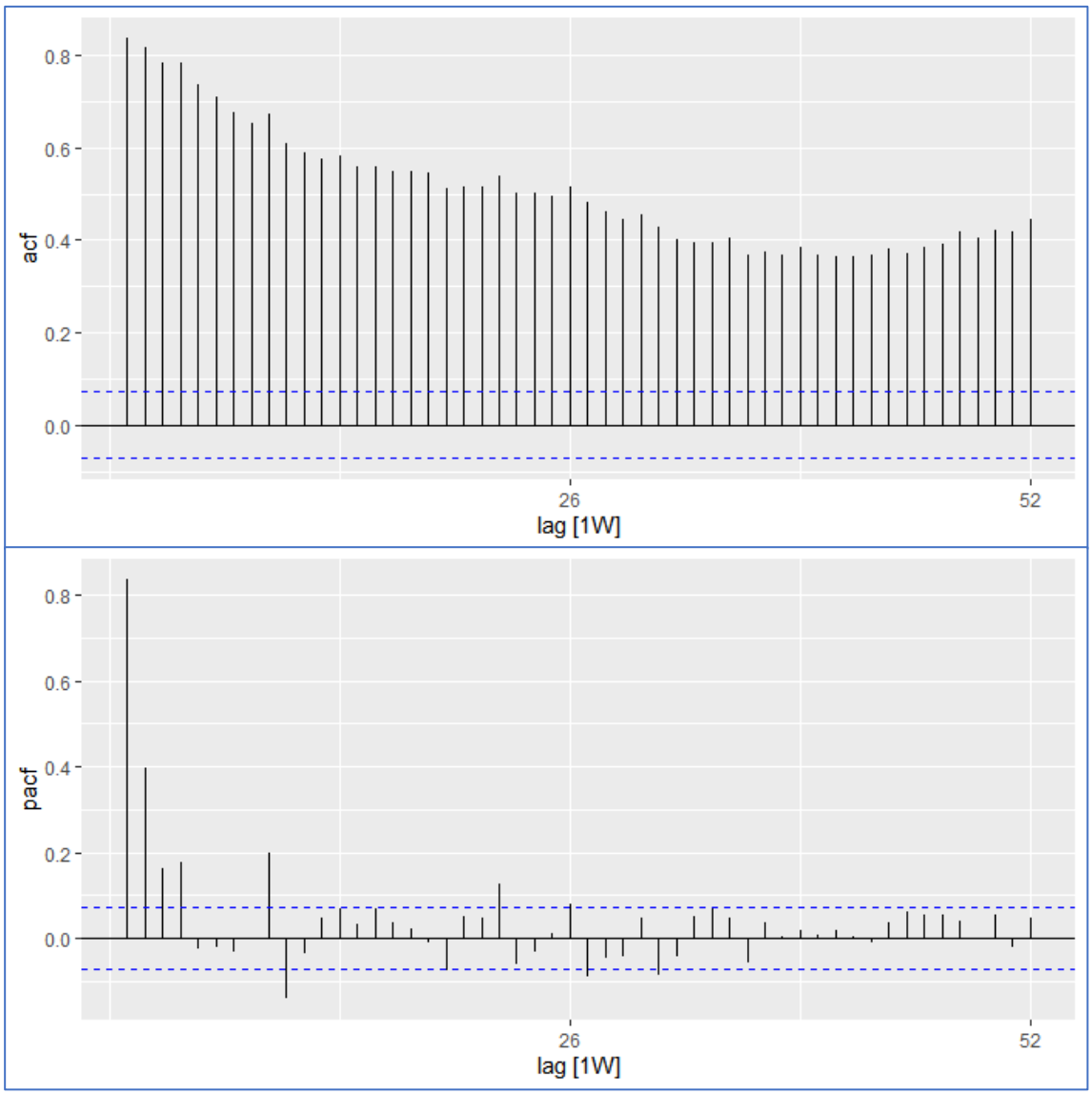
componente sazonal na série também, como podemos observar no gráfico de decomposição.



Vamos analisar a autocorrelação e a autocorrelação parcial. Queremos investigar se os dados mais recentes possuem uma correlação com seus valores no passado.

```
##autocorrelação
tssihdf %>%
  ACF(CASOS_INT, lag_max = 52) %>%
  autoplot()

##autocorrelação parcial
tssihdf %>%
  PACF(CASOS_INT, lag_max = 52) %>%
  autoplot()
```

Pela plotagem da autocorrelação e da autocorrelação parcial podemos ver que os dados apresentam uma autocorrelação com seus valores no passado, embora a autocorrelação parcial não se apresente significativa em todos os momentos.



Nós vamos realizar um teste de independência dos dados, por meio do teste de hipótese Box-Ljung para verificar se os dados são independentes entre si.

```
##teste de independencia
Box.test(tssihdf$CASOS_INT, type = "Ljung-Box")
```

Receberemos o seguinte *output* apresentando os dados testados, o valor da abcissa na distribuição X^2 (Qui-quadrado), os graus de liberdade e o *p-value*, que é menor que 0,05.

```
        Box-Ljung test

data:  tssihdf$CASOS_INT
X-squared = 512.4, df = 1, p-value < 2.2e-16
```

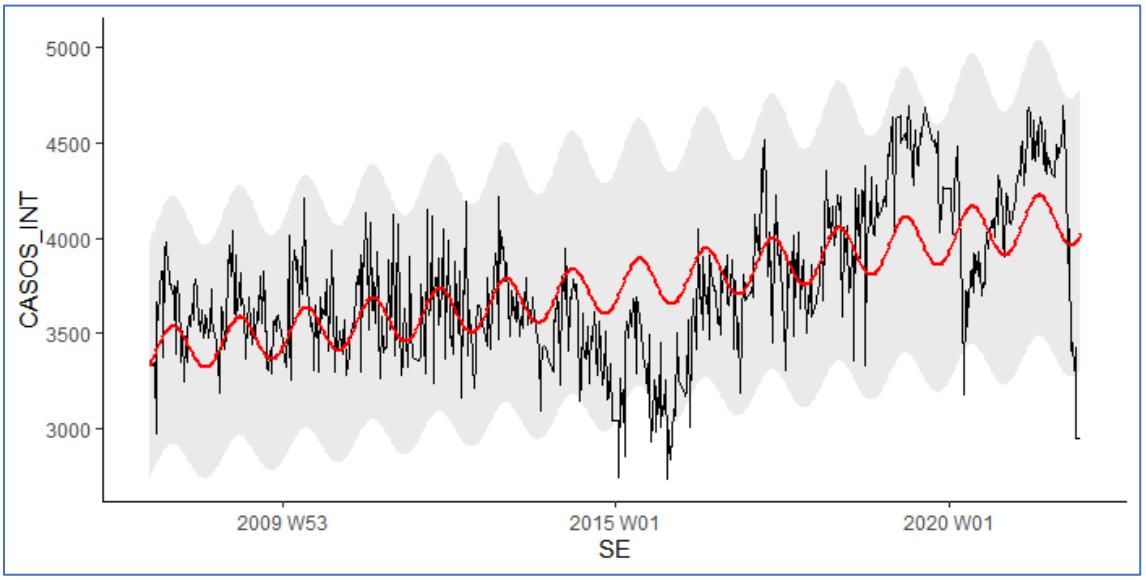
Assim, seguiremos para o ajuste da regressão binomial negativa, iniciando pela inclusão de termos Fourier com o K = 1. Após isso, vamos criar uma forma de modelo binomial negativo que incluirá a série pela semana epidemiológica + termos Fourier.

```
#ajustando uma regressão

##adicionando fourier terms usando semana epidemiológica e
casos interpolados
tssihdf$FOURIER <- tssihdf %>%
  select(CASOS, SE, CASOS_INT) %>%
  fourier(K = 1)

#modelo binomial negativo
model <- glm_nb_model(CASOS_INT ~
                        SE +
                        FOURIER)
```

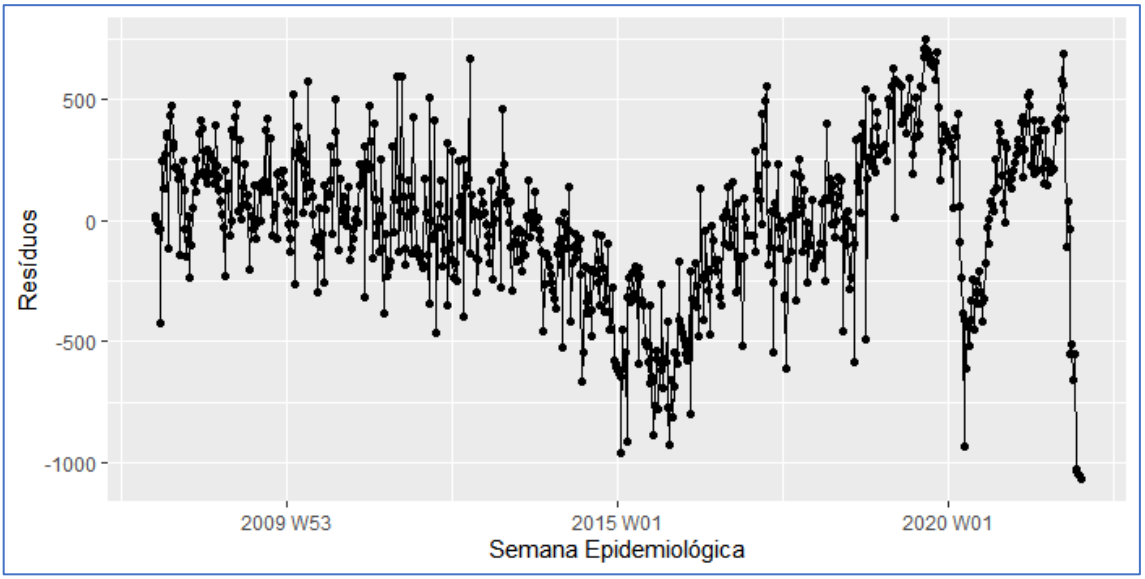
Assim, vamos transformar em *data.frame* a nossa série temporal para realizarmos o ajuste do modelo e depois vamos criar um objeto com os valores preditos para comparar com os valores observados em um gráfico.

```
##ajuste o modelo usando o dataset
tsdf = as.data.frame(tssihdf)
fitted_model <- trending::fit(model, tsdf)

## calculate confidence intervals and prediction intervals
obs <- predict(fitted_model, simulate_pi = FALSE)

## plotando o gráfico da regressão
ggplot(data = obs, aes(x = SE)) +
  geom_ribbon(aes(ymin = lower_pi,
                  ymax = upper_pi),
              alpha = 0.1) +
  geom_line(aes(y = CASOS_INT),
            col = "black") +
  geom_line(aes(y = estimate),
            col = "Red",
            size = 1) +
  theme_classic()
```

Obteremos o seguinte gráfico. A linha em vermelho representa os valores estimados e a sombra em cinza apresenta o intervalo de confiança de 95% para os valores estimados. Perceba que os valores reais, representados em preto, foram bem representados pelo modelo, sendo que a maioria ficou no intervalo de confiança estimado.



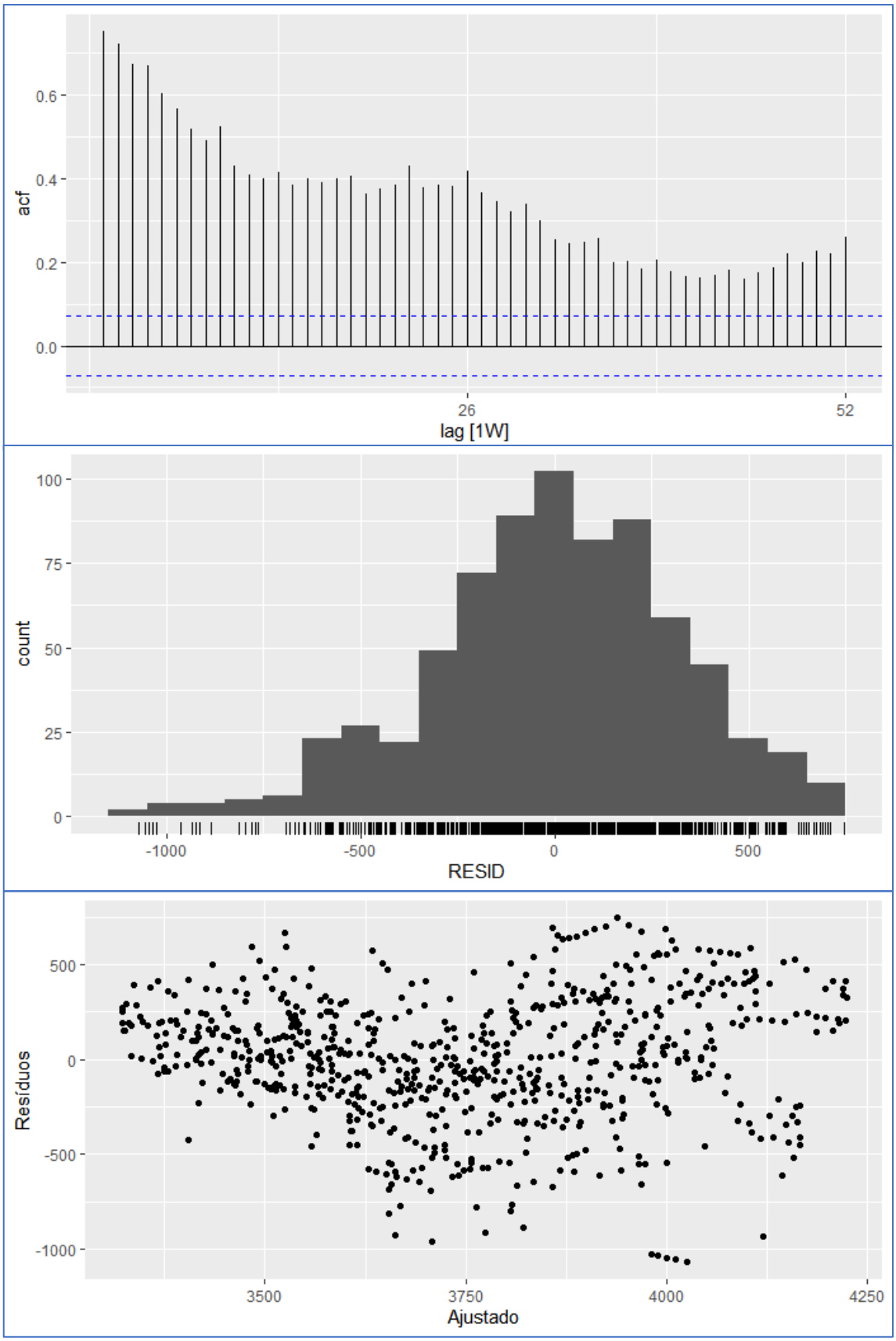
Nós analisaremos os resíduos para investigar se há algum padrão que não foi capturado pelo modelo. Analisando o gráfico do modelo, não encontramos nenhum padrão diferente, parece que há aleatoriedade no modelo.



Após isso, vamos visualizar os gráficos de autocorrelação e de autocorrelação parcial. Nesse caso, não queremos uma correlação entre os resíduos.

```
##os resíduos são autocorrelacionados?
obs %>%
  as_tsibble(index = SE) %>%
  ACF(RESID, lag_max = 52) %>%
  autoplot()

##os resíduos apresentam uma distribuição normal?
obs %>%
  ggplot(aes(x = RESID)) +
  geom_histogram(binwidth = 100) +
  geom_rug() +
  labs(y = "count")

##existe algum padrão nos resíduos?
obs %>%
  ggplot(aes(x = estimate, y = RESID)) +
  geom_point() +
  labs(x = "Ajustado", y = "Resíduos")
```

Nesse caso, o gráfico apresenta uma correlação, e uma distribuição que não parece aderir à distribuição normal padrão. Apesar disso, o gráfico de dispersão parece não apresentar padrões que possam ser investigados na série.


acf
0.6
0.4
0.2
0.0
26
52
lag [1W]
count
100
75
50
25
0
-1000
-500
0
500
RESID
Resíduos
500
0
-500
-1000
3500
3750
4000
4250
Ajustado

Por fim, realizaremos novamente o teste Box-Ljung para testar a autocorrelação dos resíduos.

```
##teste de autocorrelação dos resíduos
Box.test(obs$RESID, type = "Ljung-Box")
```

Receberemos o seguinte *output*.

```
        Box-Ljung test

data:  obs$RESID
X-squared = 412.79, df = 1, p-value < 2.2e-16
```

Referências Bibliográficas

1. MORETTIN, Pedro A.; TOLOI, Clélia. **Análise de séries temporais**. In: Análise de séries temporais. 2006. p. 538-538.
2. NIELSEEN, Alieen. **Análise Prática de Séries Temporais:** Predição com estatística e aprendizado de máquina. Alta Books, Rio de Janeiro, 2021.
3. FÁVERO, Luiz Paulo; BELFIORE, Patrícia. **Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®.** Elsevier Brasil, 2017.

Atividades

1	O conceito de série temporal não pode ser amplamente aplicado na Epidemiologia e na Vigilância em Saúde.	V	
		F	X
2	Séries Temporais podem ser definidas como um conjunto de observações ordenadas no tempo.	V	X
		F	
3	Quando temos uma variável ocorrendo de forma periódica no tempo, temos uma série temporal	V	X
		F	
4	Sistemas de informação em saúde não possuem dados que possam ser usados em análises de séries temporais.	V	
		F	X
5	Com o monitoramento ininterrupto e a coleta de dados se ampliando, aumentará a necessidade e a possibilidade de análise de séries temporais	V	X
		F	
6	Quando falamos de análises de séries temporais, estamos nos referindo a calcular a média e a mediana temporal de dados de doenças.	V	
		F	X
7	O modelo de regressão binomial negativo é enquadrado nos modelos para dados de contagem, ou seja, é aplicado aos dados com valores inteiros e não negativos.	V	X
		F	
8	O modelo de regressão binomial negativo é usado para estimar dados categóricos que mostram a presença ou ausência de uma determinada condição.	V	
		F	X
9	O modelo binomial negativo é útil para a análise de incidência de casos novos no tempo.	V	X
		F	
10	O modelo de regressão binomial negativo é baseado na distribuição binomial negativa.	V	X
		F	
11	A distribuição binomial negativa não é conhecida como Poisson-Gama.	V	
		F	X
12	Usamos o pacote *Pacman* para instalar e carregar pacotes utilizados no R.	V	X
		F	
13	Utilizamos o gráfico *boxplot* para analisar visualmente *outliers*	V	X
		F	
14	Utilizamos uma técnica de interpolação para corrigir dados defasados ou faltantes.	V	X
		F	
15	Utilizamos a decomposição para compreender os componentes das séries temporais.	V	
		F	X