Curso de Vigilância em Saúde apoiado por plataforma BI

Aula 08- Detectando surtos com análise de séries temporais

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 nossa primeira aula prática! Para completá-la é importante que você tenha o R e o RStudio instalados em seu computador. Estamos muito animados para que você participe dessa jornada com a gente!

2. Introdução

Na aula anterior, aprendemos a estimar um modelo binomial negativo utilizando os dados de internação do Distrito Federal. 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 (1).

Além de dados de internação, o modelo binomial negativo pode ser útil para analisar a incidência de doenças. Mas tenha atenção! Esse modelo é usado para valores inteiros e não negativos. Ou seja, taxas de incidência não são adequadas para um ajuste de modelo, mas apenas a contagem de seus casos em relação ao período (1).

A análise de séries temporais permite a predição de valores para o futuro, permitindo investigar o que seria esperado para o futuro dado o comportamento passado da série (2, 3). Essa propriedade pode ser útil na identificação de surtos ou pandemias.

3. Predições com séries temporais

Considerando que as observações sejam autocorrelacionadas entre si com o seu valor em períodos anteriores, podemos realizar a predição de séries temporais, isto é, estimar quais serão seus valores no futuro (2).

A predição de séries temporais univariadas permite predizer os valores da série, caso todas as condições em que o ajuste ocorreu estejam presentes (2, 3). Isso quer dizer que ela permite identificar mudanças na ocorrência dos casos, sendo que, se há valores maiores do que os estimados, podemos considerar que há um surto ou epidemia, dependendo do território em que isso foi realizado.

Nessa aula, vamos realizar a predição da série temporal que ajustamos na última aula para identificar o que ultrapassaria o número de casos esperados para o período posterior. A análise será muito semelhante ao que realizamos com o diagrama de controle, porém com uma técnica mais sofisticada.

4. Detectando surtos com análise de regressão binomial negativa

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.

Após abrir o projeto e a pasta na aula 8, nós vamos executar um comando para executar todo o *script* da aula 7, pois precisaremos dos objetos criados na aula anterior para executar nossa análise.

```
# rodando o script da aula 7
ll <- parse(file = "scripts/07_series_temporais.R")

for (i in seq_along(ll)) {
  tryCatch(eval(ll[[i]]),
           error = function(e) message("Oops!    ",
as.character(e)))
}
```

Nós vamos definir uma data para ajustar nosso modelo e uma data para comparar as nossas predições. Nós vamos realizar um corte na última semana de 2019. Como temos dados até 2021, vamos simular a identificação de um surto considerando os dados dos anos de 2020 e 2021 com as predições dos dados do modelo ajustado entre 2008 e 2019.

A ideia é predizer quantas internações seriam esperadas para 2020 e 2021. Se em algum ponto começarmos a ter mais casos do que foi estimado, podemos considerar ações para a contenção de um surto ou epidemia, a depender da extensão do território.

```
#Detecção de surtos

##definindo data de início e de fim
inicio = min(sihdf2$DATA)

##definindo a semana de corte para a data
```

```
corte = yearweek("2019-12-31")

##definindo a semana última data da série
fim = yearweek("2021-12-31")

##definindo número de semanas de interesse
nse = as.numeric(fim - corte)
```

Com as semanas de análise selecionadas, nós vamos realizar o mesmo agrupamento de casos por semana que realizamos na aula anterior. Também incluiremos os termos de Fourier para a análise entre 20089 e 2019.

```
##ajustando dados até o final do ano
tssihdf <- tssihdf %>%
  group_by_key() %>%
  group_modify(~add_row(.,
                        SE = seq(max(.$SE) + 1,
                                 fim,
                                 by = 1)))
## definindo os termos de fourier
tssihdf <- tssihdf %>%
  mutate(
    FOURIER2 = rbind(
      fourier(
        filter(tssihdf,
               SE <= corte),
        K = 1
      ),
      fourier(
        filter(tssihdf,
               SE <= corte),
        K = 1,
        h = nse
      )
    )
  )
```

Nós vamos dividir os dados para treinamento do modelo e teste. Os dados anteriores a 2019 servirão para treinar nosso modelo. Nós vamos gerar predições até 2021 e depois comparar com os dados de teste, que são as observações que efetivamente ocorrerão em 2020 e 2021.

```
##separando os dados para o treinamento e teste do modelo
```

```
dat <- tssihdf %>%
  group_by(SE <= corte) %>%
  group_split()
```

Após isso, definiremos novamente o modelo, considerando os casos como dependentes da semana epidemiológica.

```
##definindo um ajuste de modelo binomial negativo
model <- glm_nb_model(
  CASOS_INT ~
    SE +
    FOURIER2
)
```

Vamos realizar a definição também das predições, considerando os dados que usamos para o treino e teste do nosso modelo.

```
##definindo um ajuste de modelo binomial negativo
model <- glm_nb_model(
  CASOS_INT ~
    SE +
    FOURIER2
)
```

Com todas as definições ajustadas, vamos prosseguir para o ajuste do modelo. Além disso, vamos executar o intervalo de confiança e vamos salvar as predições em um *data.frame* justamente com os dados observados para que possamos consultar e comparar os dados posteriormente.

```
##definindo que data será usada para para o ajuste e
qual será usada para a predição
ajuste = pluck(dat, 2)
ajuste2 = as.data.frame(ajuste)

predicao <- pluck(dat, 1) %>%
  select(CASOS_INT, SE, FOURIER2)
predicao2 = as.data.frame(predicao)

##ajustando o modelo
modelo_ajustado <- trending::fit(model, ajuste2)

##gerando intervalo de confiança e estimativas para o
modelo ajustado
obs <- modelo_ajustado %>%
  predict(simulate_pi = FALSE)
```

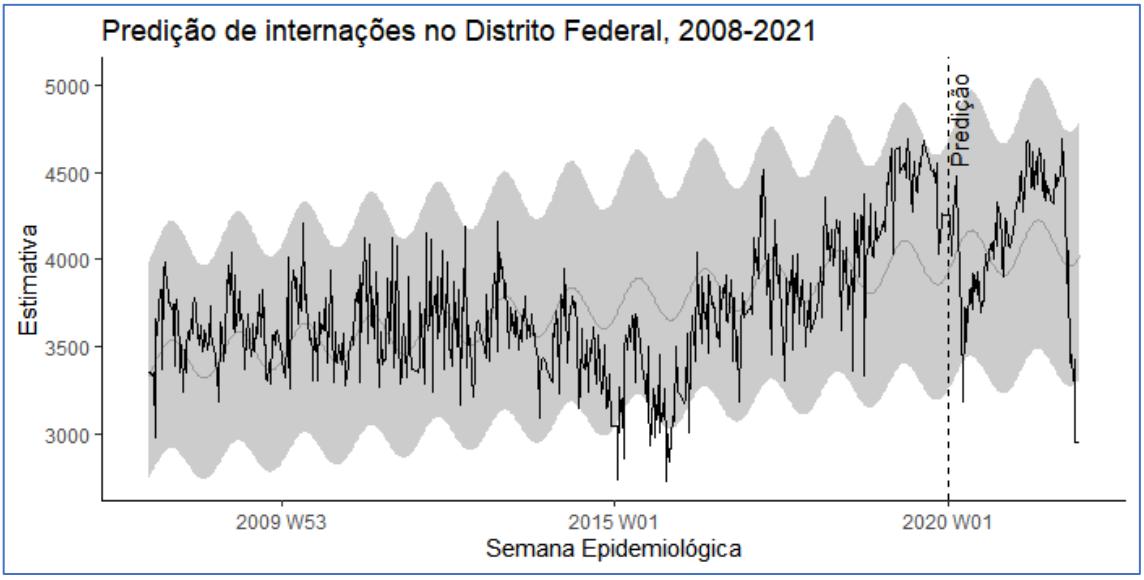
```
##gerando predições para o modelo ajustado
pred <- modelo_ajustado %>%
  predict(predicao2, simulate_pi = FALSE)

#unindo as estimativas e predicoes
obs <- bind_rows(obs, pred)
```

Nós vamos gerar o gráfico de predição até 2021 do modelo que ajustamos, juntamente com as observações e o intervalo de confiança.

```
#gerando o gráfico da regressão
ggplot(data = obs, aes(x = SE)) +
  geom_line(aes(y = estimate),
            col = "grey") +
  geom_ribbon(aes(ymin = lower_pi,
                  ymax = upper_pi),
              alpha = 0.25) +
  geom_line(aes(y = CASOS_INT),
            col = "black") +
  geom_point(
    data = filter(obs, CASOS_INT > upper_pi),
    aes(y = CASOS_INT),
    colour = "red",
    size = 2) +
  geom_vline(
    xintercept = as.Date(corte),
    linetype = "dashed") +
  annotate(geom = "text",
           label = "Predição",
           x = corte,
           y = max(obs$upper_pi) - 250,
           angle = 90,
           vjust = 1
  ) +
  theme_classic()+
  ylab("Estimativa")+xlab("Semana Epidemiológica")+
  ggtitle("Predição de internações no Distrito Federal,
2008-2021")
```

Perceba que as observações do ano de 2020 e 2021 ficaram dentro do estimado para os dois anos, com exceção de duas semanas em que as observações foram menores do que o limite inferior estimado.



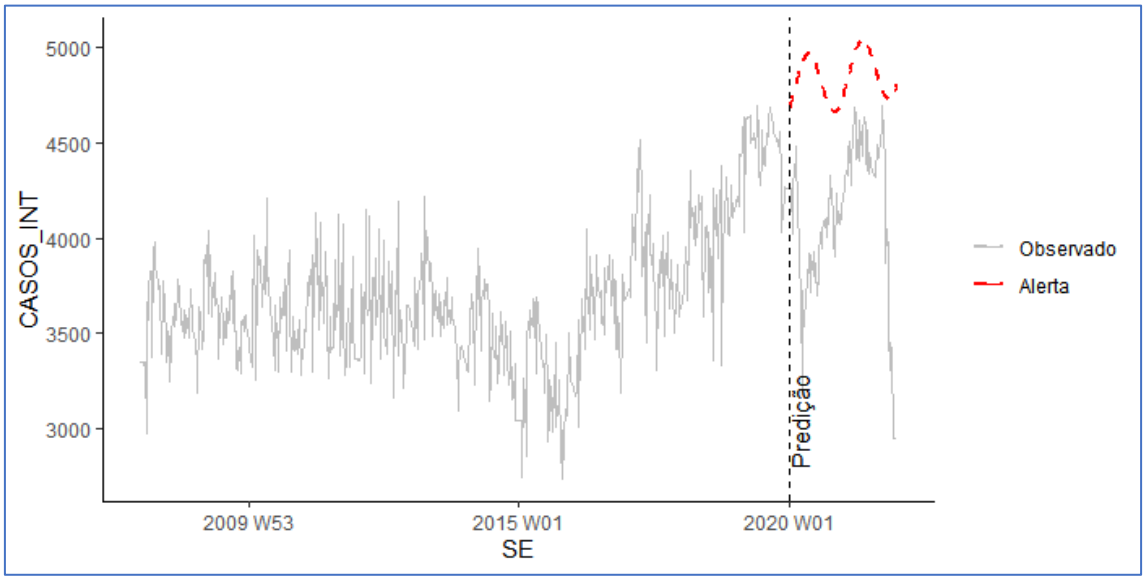
Por fim, nós vamos gerar um limiar de alerta, considerando o intervalo de confiança de 95% estimado para o período posterior à predição. Usando esse limiar, poderemos identificar valores superiores ao esperado como um alerta de possível surto.

```
#gerando um limiar de alerta
tssihdf$threshold  =  ifelse(tssihdf$SE  >=  corte,
obs$upper_pi, NA)

#gerando um gráfico com o limiar
ggplot(tssihdf, aes(x = SE)) +
  geom_line(aes(y = CASOS_INT, colour = "Observado"))
+
  geom_line(aes(y = threshold, colour = "Alerta"),
            linetype = "dashed",
            size = 1) +
  scale_colour_manual(values = c("Observado" = "grey",
                                 "Alerta" = "red")) +
  geom_vline(
    xintercept = as.Date(corte),
    linetype = "dashed") +
  annotate(geom = "text",
           label = "Predição",
           x = corte,
           y = max(obs$upper_pi) - 2000,
           angle = 90,
           vjust = 1
  )+
  theme_classic()+
  theme(legend.title = element_blank())
```

Assim, colocamos uma linha superior vermelha no gráfico de predição de internações, indicando que valores superiores ao estimado podem ser considerados como um indicativo de início de surto ou epidemia.

O gráfico mostra que embora o número de internações tenha aumentado, ele ainda está dentro da estimativa para o período e não é motivo para alerta.



Referências Bibliográficas

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

Atividades

1	O modelo de regressão binomial negativo é um modelo para dados contínuos e categoricos	V	
		F	X
2	O modelo de regressão binomial negativo é enquadrado nos modelos para dados de contagem	V	X
		F	
3	O modelo binomial negativo é aplicado aos dados com valores inteiros e não negativos	V	X
		F	
4	O modelo binomial negativo não pode ser usado para analisar a incidência de doenças.	V	
		F	X
5	O modelo binomial negativo pode ser útil para analisar a incidência de doenças ou de eventos de saúde como internações	V	X
		F	
6	A análise de séries temporais permite a predição de valores para o futuro, permitindo investigar o que seria esperado para o futuro dado o comportamento passado da série	V	X
		F	
7	A análise de séries temporais não trabalha com predições sendo responsáveis apenas por descrever a série	V	
		F	X
8	A análise de séries temporais não pode ser usada na identificação de surtos.	V	
		F	X
9	Considerando que as observações sejam autocorrelacionadas entre si com o seu valor em períodos anteriores, podemos realizar a predição de séries temporais	V	X
		F	
10	A predição de séries temporais univariadas permite predizer os	V	X
		F	

	valores da série, caso todas as condições em que o ajuste ocorreu estejam presentes		
11	As séries temporais não trabalham com o conceito de autocorrelação.	V	
		F	X
12	A estimativa de uma série temporal permite identificar mudanças na ocorrência dos casos, sendo que, se há valores maiores do que os estimados, podemos considerar que há um surto ou epidemia,	V	X
		F	
13	A análise de séries temporais é uma técnica mais sofisticada para a detecção de surtos e epidemias	V	X
		F	
14	O território deve ser considerado quando fazemos as predições utilizando séries temporais	V	X
		F	
15	Apenas diagramas de controle permitem a identificação de surtos ou epidemias	V	
		F	X