Atividade Final Regressões

A atividade deverá ser feita em duplas e enviada para o email danilosouto@gmail.com com o assunto UP2020-PTI-REG até o dia 30/11.

Aluno: Rodrigo Renie de Braga Pinto

Informe o método utilizado e justifique. A formulação da resposta faz parte da avaliação. Exibir o código utilizado no R.

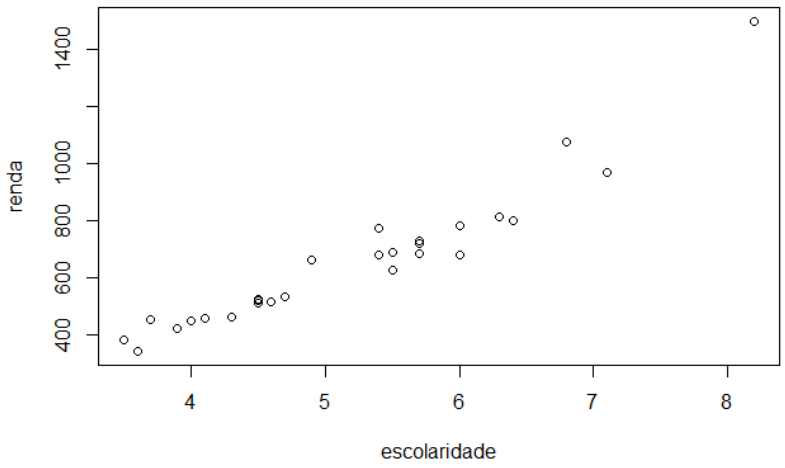
1 Faça a análise conforme descrito a seguir:

1.1 Defina a renda média *per capta* do estado em relação a média de escolaridade do estado (y = renda, x = escolaridade) em outras palavras, renda ~ escolaridade, dos dados públicos a seguir:

```
#dados para o exercício copie e cole no R

mec <- data.frame(
  row.names = c("RR", "AC", "PA", "TO", "MA", "SE", "BA", "AL", "SP", "ES",
                "SC", "PR", "GO", "DF", "AP", "RO", "AM", "PB", "RN", "PI",
                "PE","CE","RJ", "MG","RS","MT","MS"),
  escolaridade = c(5.7, 4.5, 4.7, 4.5, 3.6, 4.3, 4.1, 3.7, 6.8, 5.7, 6.3, 6.0,
                   5.5, 8.2, 6.0, 4.9, 5.5, 3.9, 4.5, 3.5, 4.6, 4.0, 7.1, 5.4,
                   6.4, 5.4, 5.7),
  renda = c(685, 526, 536, 520, 343, 462, 460, 454, 1076, 722, 814, 782, 689,
            1499, 683, 662, 627, 423, 513, 383, 517, 448, 970, 681, 800, 775,
            731)
)
```

1.2 Veja os gráficos de dispersão:

```
plot(mec)
```

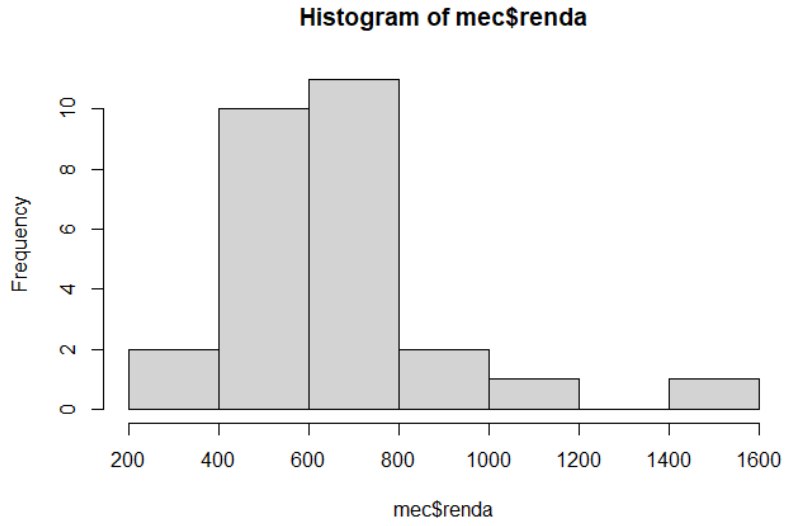


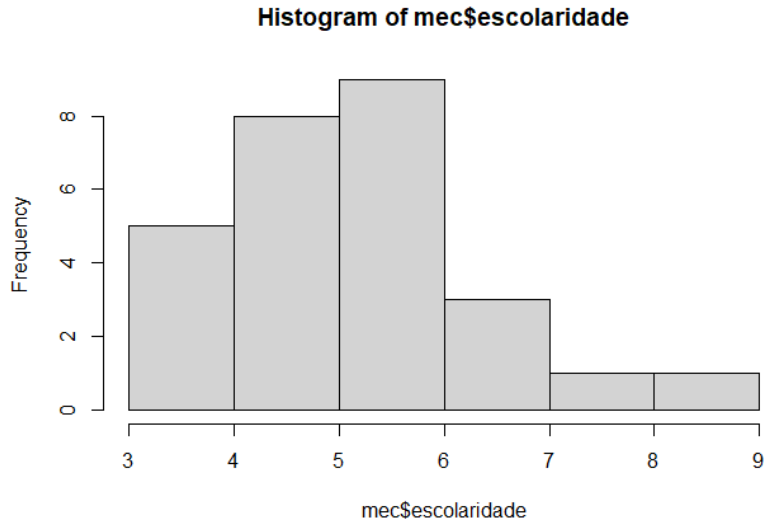
1.3 Exiba as correlações:

```
cor(mec)
             escolaridade     renda
escolaridade    1.0000000 0.9507579
renda           0.9507579 1.0000000
```

1.4 Plote os histogramas de renda e escolaridade:

```
hist(mec$renda)
hist(mec$escolaridade)
```



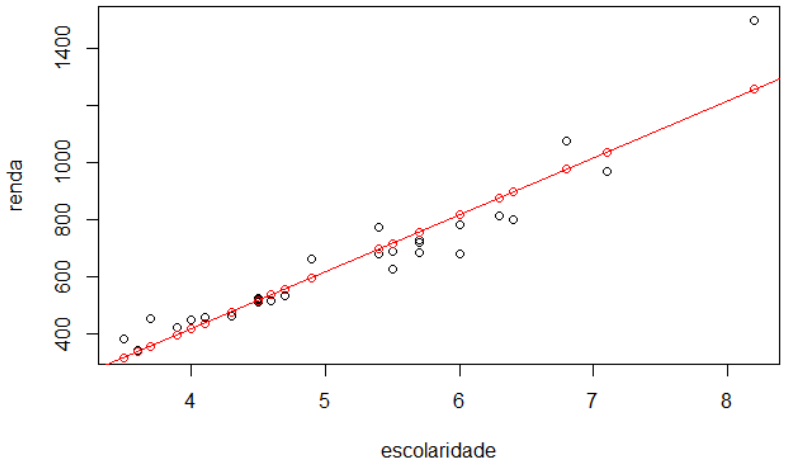


1.5 Teste de normalidade:

```
# Shapiro Test
# H0: x  = normal (há normalidade)
# H1: x != normal (não há normalidade)

if (shapiro.test(mec$renda)$p.value > 0.05) {
  print('Aceita-se a Hipósete Nula (H0)')
} else {
  print('Rejeita-se a Hipósete Nula (H0)')
}

if (shapiro.test(mec$escolaridade)$p.value > 0.05) {
  print('Aceita-se a Hipósete Nula (H0)')
} else {
  print('Rejeita-se a Hipósete Nula (H0)')
}

# mec$renda não segue uma distribuição normal
# mec$escolaridade segue uma distribuição normal
```

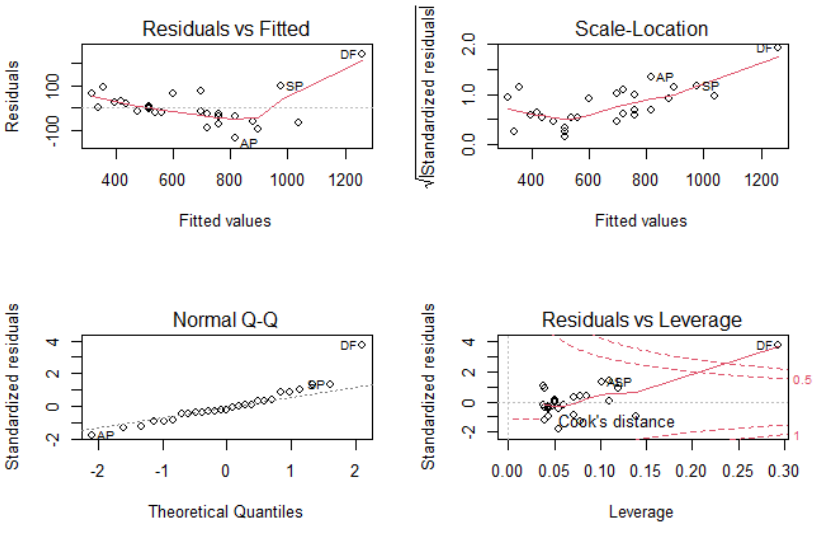
1.6 Faça a regressão linear *lm():*

```
modelo.linear <- lm(renda ~ escolaridade, data=mec)

plot(mec)
abline(modelo.linear, col='red')
points(mec$escolaridade, modelo.linear$fitted.values, col='red')
```



1.7 Quais são os pontos com maior alavancagem?

```
sort(influence(modelo.linear)$hat)
layout(matrix(c(1, 2, 3, 4), 2, 2))
plot(modelo.linear)
```

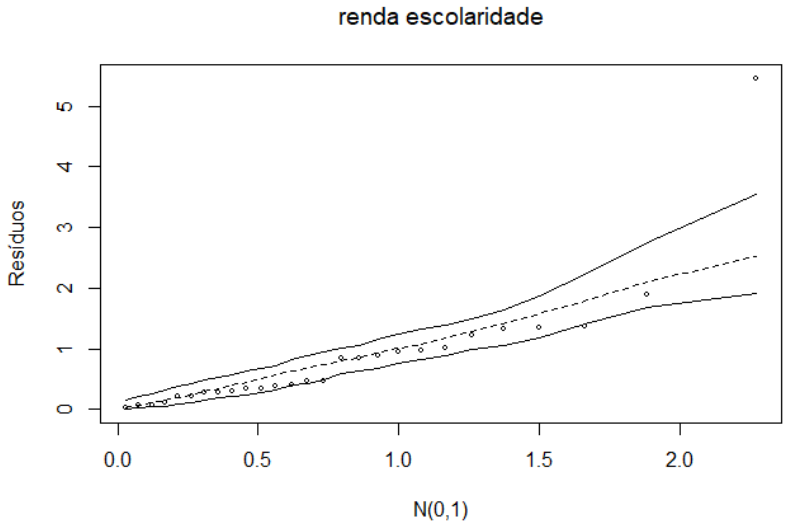


1.8 Qual o coeficiente de determinação (R-squared)?

```
summary(modelo.linear)$r.squared
```

1.9 Verifique os resíduos com a biblioteca *library(hnp):*

```
library(hnp)
hnp(modelo.linear, xlab='N(0,1)', ylab='Resíduos',
    main=modelo.linear$call$formula)
```



1.10 A regressão linear parece ser uma boa escolha? Por que?

Não, pois um dos principais requisitos do modelo linear é que as variáveis preditoras sigam uma distribuição normal, o que não ocorre com a escolaridade, como pode ser visto no item 1.5. Isso se reflete na análise dos resíduos deste modelo do item 1.9, onde os pontos se afastam da linha central e ainda extrapolam os limites do gráfico.

1.11 Qual das distribuições que estudamos (Binomial, Normal, Poisson, Gama, Gaussiana Inversa) tem uma semelhança com os dados mostrados pelo histograma de renda?

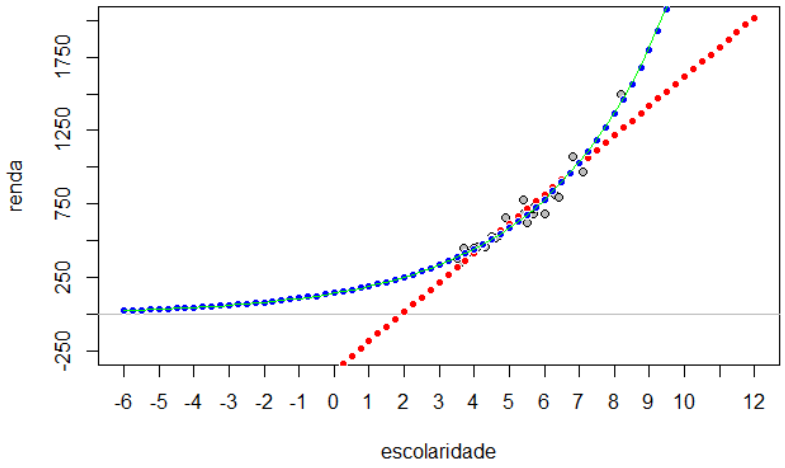
Tanto a Gama quando a Gaussiana Inversa tem semelhanças significativas com o histograma de renda. Qual das duas distribuições será de fato melhor deverá ser analisado a seguir.

1.12 Faça uma regressão *glm()* com essa distribuição:

```
modelo.gamma <- glm(renda ~ escolaridade, data=mec, family=Gamma(link='log'))
modelo.invgaus <- glm(renda ~ escolaridade, data=mec,
                      family=inverse.gaussian(link='log'))
```

1.13 Agora estime (utilize a função *predict()*) os valores de renda para os valores de escolaridade utilizando os dois modelos (*lm()* e *glm()*) e plote os gráficos com as curvas. Mostre no mesmo gráfico os valores observados em preto, preditos do modelo 1 em vermelho, e preditos no modelo 2 em verde (utilize as funçoes *plot(), points() e points()*):

```
predict.values <- seq(from=-6, to=12, by=0.25)
predict.dataframe <- data.frame(escolaridade=predict.values)

predict.linear <- predict(modelo.linear, predict.dataframe, type='response')
predict.gamma <- predict(modelo.gamma, predict.dataframe, type='response')
predict.invgaus <- predict(modelo.invgaus, predict.dataframe, type='response')

plot(mec, data=mec, ylim=c(-250,2000), xlim=c(-6,12), col='black', bg='grey',
     pch=21, xaxt='n', yaxt='n')
axis(side=1, at=seq(-6, 12, by=1), labels=T)
axis(side=2, at=seq(-250, 2000, by=250), labels=T)
abline(h=0, col='grey')

points(predict.values, predict.linear, pch=21, col='white', bg='red')
points(predict.values, predict.gamma, pch=21, col='white', bg='blue')
lines(predict.values, predict.invgaus, pch=21, col='green')
```



1.14 Compare os modelos com a função *AIC()* e informe qual modelo você escolhe e por que:

```
AIC(modelo.linear, modelo.gamma, modelo.invgaus)

               df      AIC
modelo.linear   3 315.2632
modelo.gamma    3 288.1337
modelo.invgaus  3 288.9597
```

Baseando-se no valor do *AIC* de cada um dos modelos, por uma diferença muito pequena, a distribuição *Gamma* é a mais apropriada.

2 Encontre uma base de dados de sua preferência, caso não possua alguma há várias disponíveis no https://www.kaggle.com/datasets e http://dados.gov.br/dataset, e faça uma análise de regressão ou *forecast* sobre alguma informação que lhe pareça importante. Atenção que todas as análises dos resultados e gráficos devem ser exibidas, comentadas e descritas abaixo.

Para este exercício, foram coletadas informações de uma base de dados de interna da Itaipu Binacional. Esta base de dados é da ferramenta de ITSM (Gerenciamento de Serviços), utilizada para registrar as solicitações de serviços da Central de Atendimento e também os incidentes que ocorrem na infraestrutura da empresa.

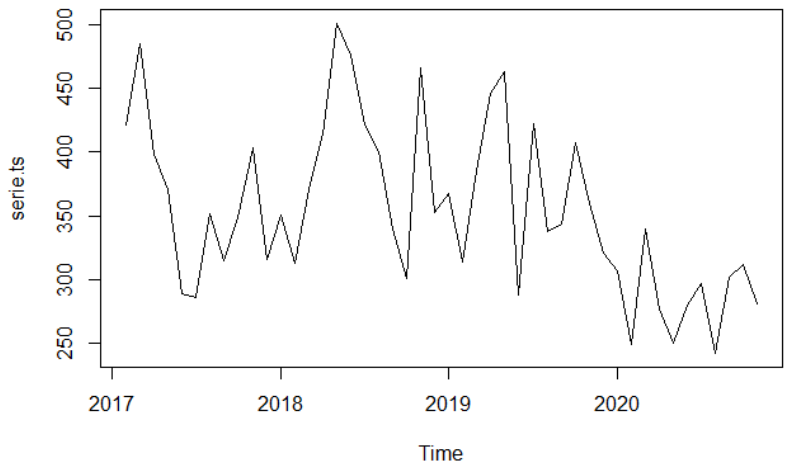
A informação coletada diz respeito ao número de incidentes mensal registrado na ferramenta, desde o mês de fevereiro de 2017 até novembro de 2020. Este período foi escolhido pois foi o processo de Gerenciamento de Incidentes da Itaipu Binacional foi efetivamente implementado na ferramenta em fevereiro de 2017. Os dados coletados estão contidas no arquivo "*dados_tickets.csv*" em anexo.

O objetivo é descobrir qual é a tendência do total de incidentes registrados para os próximos 13 meses, ou seja, até dezembro de 2021.

Primeiramente, os dados do arquivo CSV são importados no R e transformados em um *TimeSeries*:

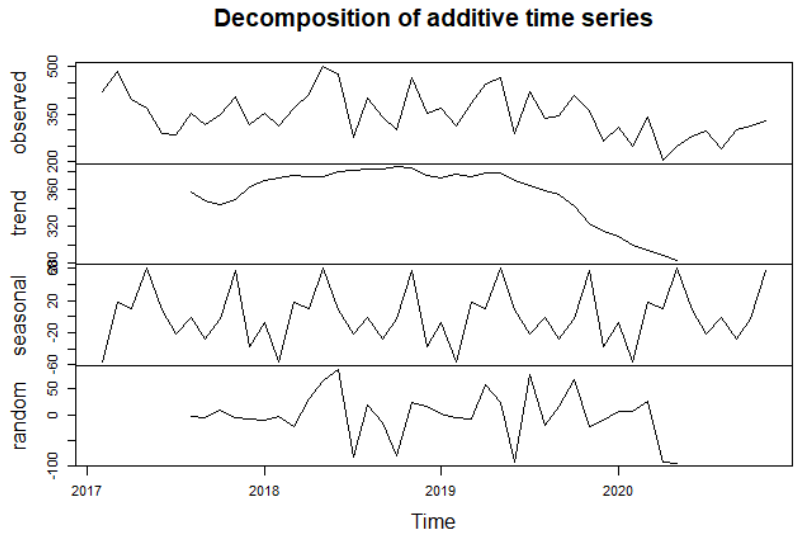
```
library(TTR)
library(forecast)
serie.dados <- read.csv('dados_tickets.csv')
serie.ts <- ts(serie.dados$TOTAL, start=c(2017, 2), end=c(2020, 11), frequency=12)
serie.ts
##      Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
## 2017     421 485 398 370 289 286 352 315 349 403 316
## 2018 351 313 370 414 501 476 421 400 340 301 466 353
## 2019 367 314 385 446 463 288 422 338 344 407 359 321
## 2020 307 249 340 277 250 280 297 242 302 312 281
plot(serie.ts)
```

A frequência 12 é utilizada na função *ts()* por se tratar de dados coletados mensalmente. Observando o gráfico resultante percebe-se, a princípio, uma queda no total de incidentes registrados:



Esta observação pode ser confirmada ao se analisar a decomposição desta *TimeSeries*, conforme indicado pelo campo *trend*:

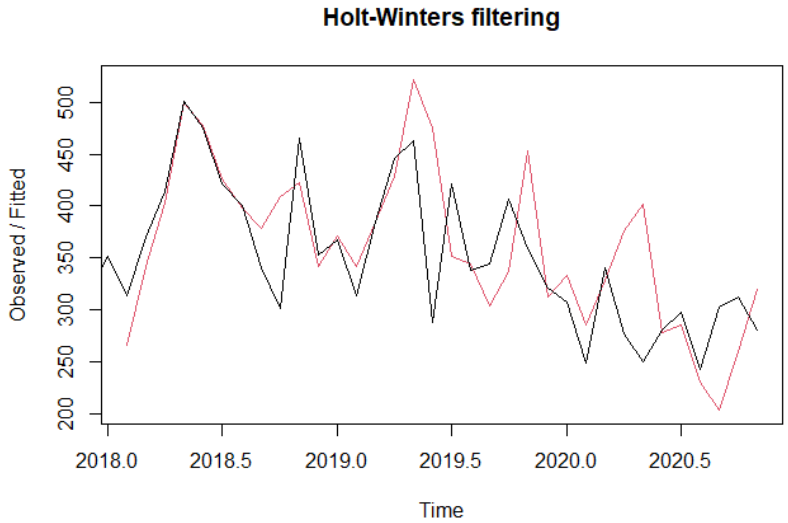
```
serie.dc <- decompose(serie.ts)
plot(serie.dc)
```



Para realizar a previsão dos dados dos próximos 6 meses, é utilizado o algoritmo Holt-Winters, da seguinte maneira:

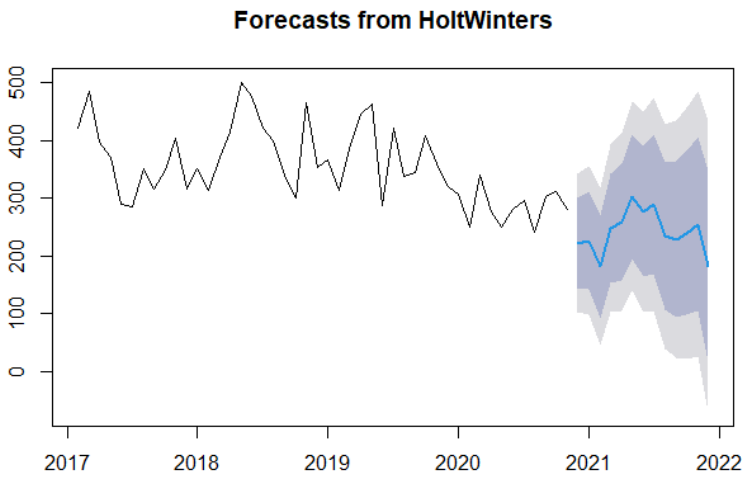
```
serie.holt <- HoltWinters(serie.ts)
plot(serie.holt)
```

Percebe-se que o algoritmo consegue se adaptar bem aos dados:



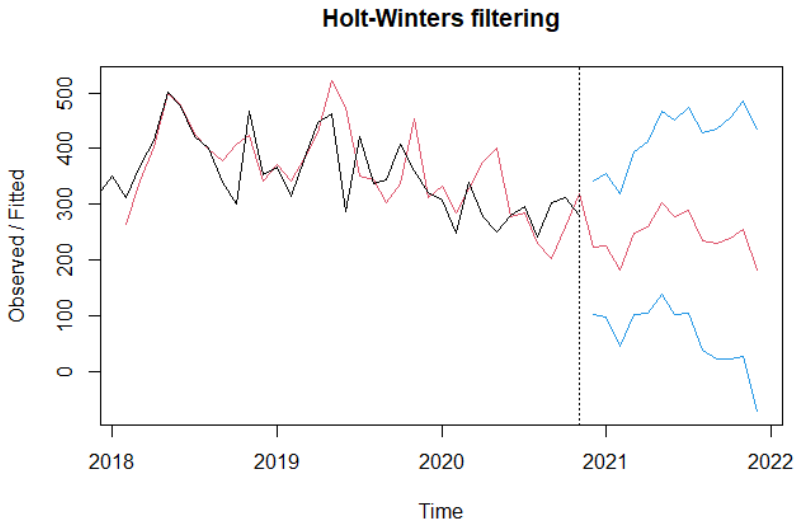
Finalmente, para realizar a previsão da quantidade de incidentes dos próximos 6 meses, utiliza-se a função *forecast()*:

```
serie.previsao1 <- forecast(serie.holt, h=13)
plot(serie.previsao1)
```



Para fins de comparação, o gráfico de previsão abaixo é gerado utilizando a função predict(), que como esperado gera dados de previsão bem semelhantes ao método anterior:

```
serie.previsao2 <- predict(serie.holt, n.ahead=13,
                           prediction.interval=T, level=0.95)
plot(serie.holt, serie.previsao2)
```



Conclusão: há evidências que a quantidade de incidentes registradas na ferramenta diminuirá ao longo de 2021, seguindo a mesma tendência de queda dos últimos 3 anos.