Estudo de covid-19: efeito da testagem no número de casos detectados

Antonio Rodrigues Neto

15/08/2021

1 Dados Brasileiros

Este estudo visa identificar se o número de casos positivos de covid detectados em uma certa localidade é efetado pela quantidade de testes realizados na mesma locidade e no mesmo período. Foram escolhidos dados mensais observados nos estados do Brasil para realizar o estudo. Este constructo será avaliado por meio da técnica supervisionada de Regressão Simples.

Os dados do brasil foram obtidos do seguinte diretório no GitHub: https://github.com/wcota/covid19br/. A importação dos dados no R é realizada da seguinte forma:

```
general <- read.csv(file = 'dada_states.csv',sep = ";")
general$date <- as.Date(general$date, format= "%d/%m/%Y")
```

Os dados de testes, casos e mortes acumulados do dataset *general* são utilizados para obter números mensais dos primeiros 7 meses de 2021, conforme:

```
months_i<-c("2021-01-01","2021-02-01","2021-03-01","2021-04-01","2021-05-01",
            "2021-06-01","2021-07-01")
```

```
months_f<-c("2021-01-31","2021-02-28","2021-03-31","2021-04-30","2021-05-31",
            "2021-06-30","2021-07-31")
cols <- c("state","date","deaths_per_100k_inhabitants",
          "totalCases_per_100k_inhabitants",
          "vaccinated_per_100_inhabitants",
          "tests_per_100k_inhabitants")
dados_mes<-list()
for (i in 1:(length(months_i)-1)){
  first <- general[,cols] %>% filter(general$date == as.Date(months_i[i]))
  last <- general[,cols] %>% filter(general$date == as.Date(months_f[i]))
  merged <- merge(first,last,by.x = "state", by.y="state")
  merged$month<-month.abb[month(months_i[i])]
  dados <- merged[,c("state","month")]
  dados["monthly_deaths_100k"]<-merged$deaths_per_100k_inhabitants.y-
    merged$deaths_per_100k_inhabitants.x
  dados["monthly_cases_100k"]<-merged$totalCases_per_100k_inhabitants.y-
    merged$totalCases_per_100k_inhabitants.x
  dados["monthly_tests_100k"]<-merged$tests_per_100k_inhabitants.y-
    merged$tests_per_100k_inhabitants.x
  dados<- filter(dados,dados$monthly_tests_100k !=0)
  dados<- dados[-c(nrow(dados)),]
  dados_mes[[i]]<-dados
}
dados1<-do.call("rbind",dados_mes)
```

Assim, o dataset *dados1* contém as informações desejadas. Vale citar que não foi necessário agrupar os dados do dataset original, visto que os valores acumulados já estavam classificados

por estados do Brasil. Além disso, os dados originais já estavam normalizados e apresentados como casos e testes por 100 mil habitantes.

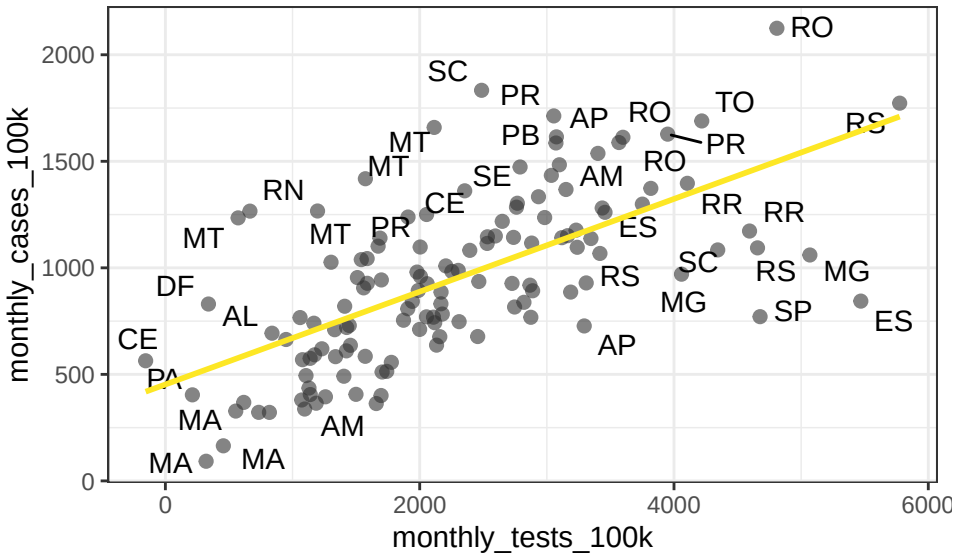
Foram retirados os outliers (5% de cada extremidade) em testagem mensal, para que a regressão não seja afetada por estes valores:

```
# Retirando outliers em testes
Q <- quantile(dados1$monthly_tests_100k, probs=c(0.05, 0.95), na.rm = FALSE)
iqr <- IQR(dados1$monthly_tests_100k)
up <-  Q[2]+1.5*iqr # Upper Range
low<- Q[1]-1.5*iqr # Lower Range
dados1<- subset(dados1, (dados1$monthly_tests_100k>low & dados1$monthly_tests_100k<up))
```

Já é possível entao visualizar o gráfico de dispersão dos dados de casos e testagem mensal, juntamente com a linha de tendências linear obtida pela função *geom_smooth* com *method = "lm"*, conforme abaixo:

```
dados1 %>%
  ggplot(aes(x = monthly_tests_100k,
             y = monthly_cases_100k)) +
  geom_point(color = "grey20", alpha = 0.6, size = 2) +
  geom_text_repel(aes(label = state)) +
  geom_smooth(aes(x = monthly_tests_100k,
                  y = monthly_cases_100k),
              method = "lm", color = "#FDE725FF", se = F) +
  theme_bw()
```

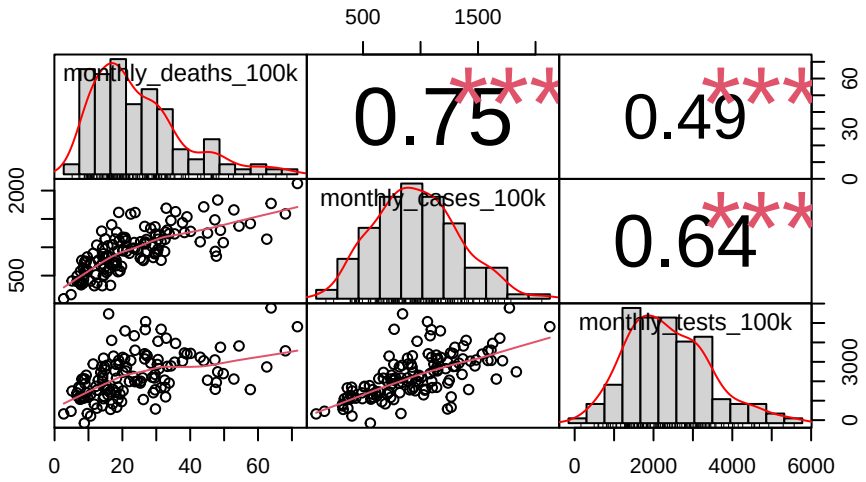
```
## `geom_smooth()` using formula 'y ~ x'
```



É possível observar que a linha de tendência no gráfico de dispersão já indica um resultado positivo para a existência de um modelo de regressão linear simples. Além disso, pode-se verificar a existência de correlação de Pearson entre as duas variáveis:

```
# Correlacao de pearson

chart.Correlation((dados1[3:5]), histogram = TRUE)
```



Aqui podemos observar a existência de uma correlação de 0.64 entre as variáveis *monthly_cases_100k* e *monthly_tests_100k*, o que indica que o modelo de regressão simples

deve obter um R^2 de aproximadamente $0.64^2 = 0.41$.

Assim, pode-se finalmente construir o modelo de Regressão Linear Simples:

```
modelo_cases <- lm(formula = monthly_cases_100k ~ monthly_tests_100k,
                       data = dados1)
summary(modelo_cases)
```

```
##
## Call:
## lm(formula = monthly_cases_100k ~ monthly_tests_100k, data = dados1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -799.52 -229.81  -14.07  176.91  839.23
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        452.62834   57.78550   7.833 1.48e-12 ***
## monthly_tests_100k   0.21771    0.02274   9.574  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 299.6 on 130 degrees of freedom
## Multiple R-squared:  0.4135, Adjusted R-squared:  0.409
## F-statistic: 91.67 on 1 and 130 DF,  p-value: < 2.2e-16
```

Conforme esperado, o modelo de regressão resulta em um $R^2 = 0.41$, signficando que aproximadamente 40% do comportamento da variável *monthly_cases_100k* é devido à variação de *monthly_tests_100k*. Além disso, o testes F para o modelo de regressão resultou

em $p - value = 2.2e - 16$, bem abaixo de um nível de significancia de 5%. Pode-se observar também que ambos os p-value obtidos para os testes T dos parametros do modelo resultaram em valores abaixo de 5%. Assim, pode-se atestar que o modelo tem signficância estatística confirmada.

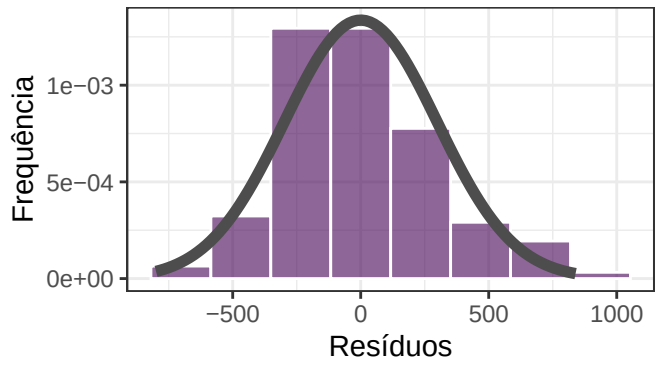
Pode-se também avaliar a qualidade do modelo verificando se os resíudos seguem a normalidade com uso do teste Shapiro-Francia:

```
# Residuos seguem normalidade?
sf.test(modelo_cases$residuals)
```

```
##
##  Shapiro-Francia normality test
##
## data:  modelo_cases$residuals
## W = 0.98316, p-value = 0.09289
```

```
dados1 %>%
  mutate(residuos = modelo_cases$residuals) %>%
  ggplot(aes(x = residuos)) +
  geom_histogram(aes(y = ..density..),
                 color = "white",
                 fill = "#440154FF",
                 bins = 8,
                 alpha = 0.6) +
  stat_function(fun = dnorm,
                args = list(mean = mean(modelo_cases$residuals),
                            sd = sd(modelo_cases$residuals)),
                size = 2, color = "grey30") +
  scale_color_manual(values = "grey50") +
```

```
  labs(x = "Resíduos",
       y = "Frequência") +
  theme_bw()
```

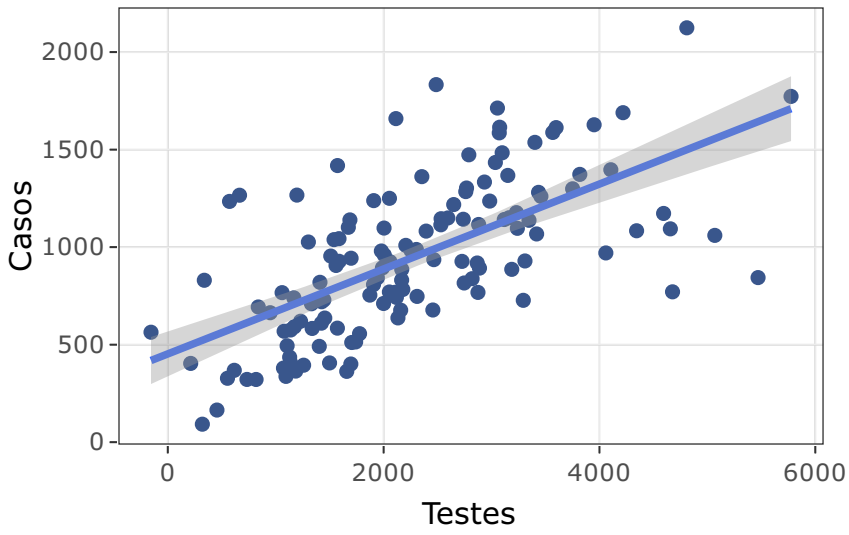


Os resultados inidicam que os resíduos seguem a normalidade. Vale lembrar que, no teste de Shapiro-Francia, existe uma inversão da hipótese nula e alternativa em relação aos demais testes, portanto busca-se um p-valor maior que o nível de confiança para atestar a normalidade.

Pode-se então visualizar o resultado do modelo de regressão:

```
# Printando modelo regressao linear
ggplotly(
  ggplot(dados1, aes(x = monthly_tests_100k, y = monthly_cases_100k)) +
    geom_point(color = "#39568CFF") +
    geom_smooth(aes(x = monthly_tests_100k, y = monthly_cases_100k),
                method = "lm",
                level = 0.95,) +
    labs(x = "Testes",
         y = "Casos") +
    theme_bw()
)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



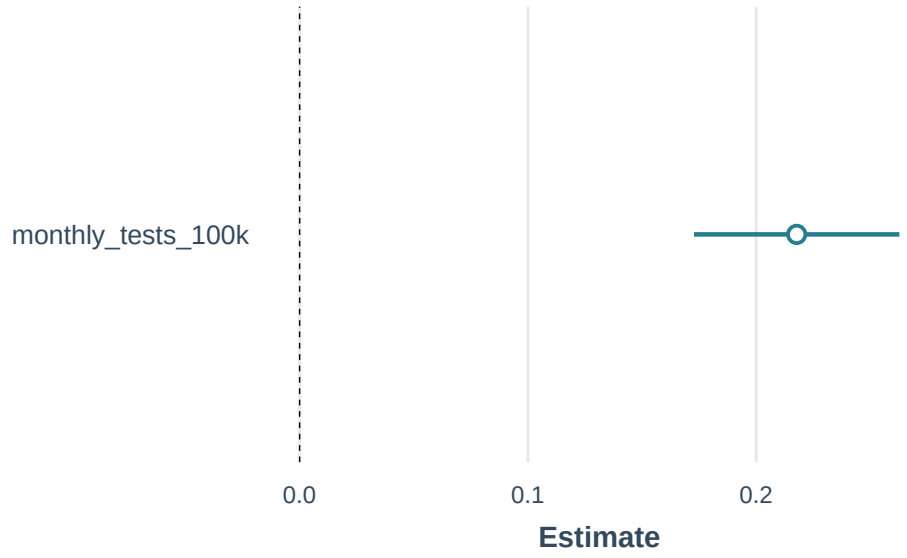
A influência da variável X na Y pode ser efetivamente avaliada por meio do parametro β da Regressão Linear Simples, conforme:

```
confint(modelo_cases, level = 0.95) # siginificância 5%
```

```
##                           2.5 %      97.5 %
## (Intercept)         338.3066305 566.9500405
## monthly_tests_100k    0.1727283   0.2627017
```

```
plot_summs(modelo_cases, colors = "#287D8EFF") #função plot_summs do pacote ggstance
```

```
## Loading required namespace: broom.mixed
```



Pode-se observar que o parametro β de *monthly_cases_100k* apresenta intervalo de confiança $[0.1727283, 0.2627017]$. Portanto, a variação positiva na testagem resulta também em uma variação positiva na identificação dos casos de covid. Em média, o aumento de 1000 testes por 100 mil habitantes resulta em um aumento de 218 casos por 100 mil habitantes.

2 Dados Europeus

Nesta segunda seção, a mesma técnica será aplicada a outro banco de dados, considerando dados europeus de cosos e testagem. A granularidade será mantida em número de casos e testes (normalizados por 100 mil habitantes) observados em um período de 1 mês durante o ano de 2021. Os dados europeus podem ser encontrados no seguinte link: https://www.ecdc.europa.eu/en/publications-data/data-virus-variants-covid-19-eueea.

```
general_ue <- read.csv(file = 'data_ue.csv',sep = ",")
general_ue$year_week<-week2date(general_ue$year_week)
nations_ue <- filter(general_ue, general_ue$level=="national")
nations_ue_2021<-filter(nations_ue,year(nations_ue$year_week)==2021)
```

Os dados originais se apresentam por semana (Wij) do ano na variável *year_week*. Estes

valores já foram transformados em dias (considerando o primeiro dia de cada semana) através da função *week2date*. O agrumento dos dados por mês e por país da Europa é realizado da seguinte forma:

```
bymonth <- nations_ue_2021 %>%
  group_by(country_code,month(year_week)) %>%
  summarize(cases = sum(new_cases, na.rm=TRUE),
            tests = sum(tests_done, na.rm=TRUE))
dados_ue <- merge(bymonth,unique(nations_ue_2021[,c("population","country_code")]),
                  by.x = "country_code", by.y="country_code")
dados_ue$cases_100k <- 100000*dados_ue$cases/dados_ue$population
dados_ue$tests_100k <- 100000*dados_ue$tests/dados_ue$population
dados_ue$`month(year_week)`<-month.abb[dados_ue$`month(year_week)`]
dados_ue$cases<-NULL
dados_ue$population<-NULL
dados_ue$tests<-NULL
# Excluindo Aug (ainda não finalizado)
dados_ue <- filter(dados_ue,dados_ue$`month(year_week)`!="Aug")
```

Vale observar que o agrupamento em países da Europa é adequado para comparação com os dados brasileiros, os quais são agrupados em estados. Pois o Brasil pode ser considerado um país continental e a população de cada estado se assemelha bastante com a população de diversos países da Europa.

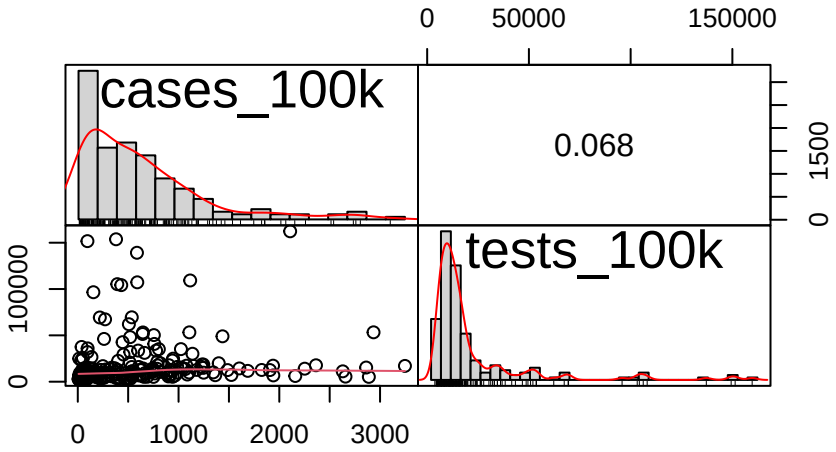
Novamente, foram retirados os outliers em testagem mensal, conforme:

```
# Retirando outliers em testes
Q_ue <- quantile(dados_ue$tests_100k, probs=c(0.05, 0.95), na.rm = FALSE)
iqr_ue <- IQR(dados_ue$tests_100k)
up_ue <-  Q_ue[2]+1.5*iqr_ue # Upper Range
```

```
low_ue<- Q_ue[1]-1.5*iqr_ue # Lower Range
dados_ue<- subset(dados_ue,
                  (dados_ue$tests_100k>low_ue & dados_ue$tests_100k<up_ue)) #removendo
```

Assim, pode-se iniciar a análise dos dados europeus. Primeiramente, a correlação entre as variáveis *cases_100k* e *test_100k* é igual a 0.068, conforme código abaixo. Esta baixa correlação indica que dificilmente um modelo de regressão que identifique influência de uma variável sobre a outra poderá ser encontrado.

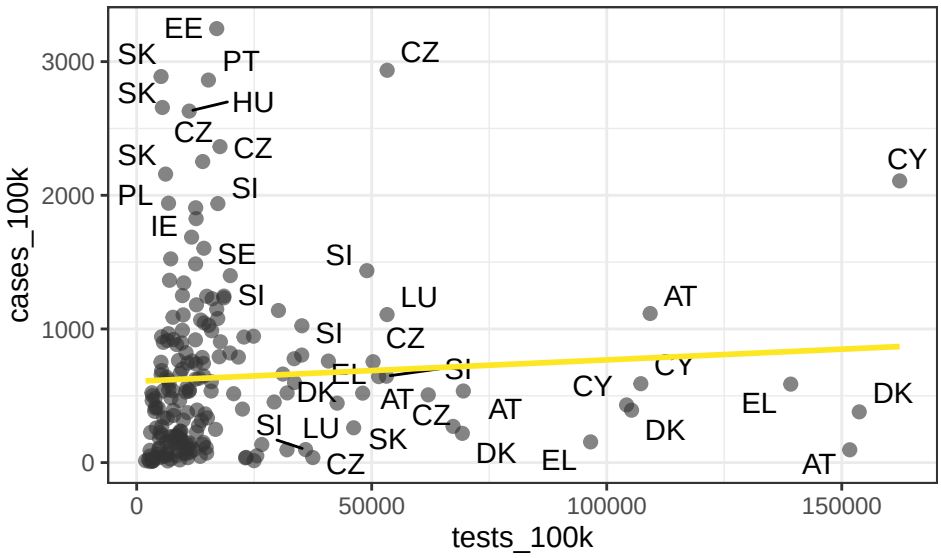
```
chart.Correlation((dados_ue[3:4]), histogram = TRUE)
```



Além disso, o gráfico de dispersão juntamente com a linha de tendências também já pode ser visualizado:

```
# GRAFICO DE DISPERSAO
dados_ue %>%
  ggplot(aes(x = tests_100k,
             y = cases_100k)) +
  geom_point(color = "grey20", alpha = 0.6, size = 2) +
  geom_text_repel(aes(label = country_code)) +
```

```
  geom_smooth(aes(x = tests_100k,
                  y = cases_100k),
              method = "lm", color = "#FDE725FF", se = F) +
  theme_bw()
```



Observa-se que a linha de tendências tem um comportamento quase veritical, o que também indica uma baixa chance de obter um modelo com significância estatística entre as duas variáveis. Então, o modelo de regressão pode ser finalmente construído:

```
# Fazendo regressao
modelo_ue <- lm(formula = cases_100k ~ tests_100k,
                data = dados_ue)
summary(modelo_ue)
```

```
##
## Call:
## lm(formula = cases_100k ~ tests_100k, data = dados_ue)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -755.5 -475.0 -162.5  273.7 2611.3
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.085e+02  5.692e+01   10.69   <2e-16 ***
## tests_100k  1.600e-03  1.685e-03    0.95    0.343
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 640.1 on 194 degrees of freedom
## Multiple R-squared:  0.004627,   Adjusted R-squared:  -0.000504
## F-statistic: 0.9018 on 1 and 194 DF,  p-value: 0.3435
```

Os resultados acima indicam que o modelo não tem significância estatística. O valor obtido de R^2 é extremamente baixo e o teste F do modelo resultou em um p-valor acima do nível de significância, bem como teste T para o parâmetro β de *tests_100k*. Estes resultados indicam que não é possível atestar uma dependência linear entre a testagem e o número de casos positivos para este dataset.

Porém, é possível ainda buscar por relações entre as variáveis com caráter não linear. Por exemplo, ao utilizar a técnica de Box-Cox. Inicialmente, verifica-se que os resíduos obtidos do modelo linear seguem a normalidade:

```
sf.test(modelo_ue$residuals)
```

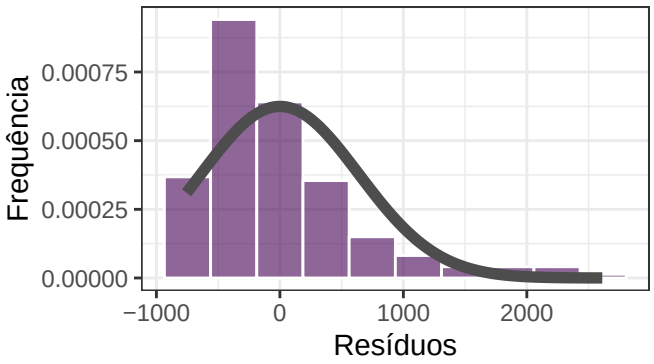
```
##
##  Shapiro-Francia normality test
##
## data:  modelo_ue$residuals
```

```
## W = 0.82386, p-value = 2.698e-12
```

```
sf.test(modelo_ue$residuals)
```

```
##
##  Shapiro-Francia normality test
##
## data:  modelo_ue$residuals
## W = 0.82386, p-value = 2.698e-12
```

```
dados_ue %>%
  mutate(residuos = modelo_ue$residuals) %>%
  ggplot(aes(x = residuos)) +
  geom_histogram(aes(y = ..density..),
                 color = "white",
                 fill = "#440154FF",
                 bins = 10,
                 alpha = 0.6) +
  stat_function(fun = dnorm,
                args = list(mean = mean(modelo_ue$residuals),
                            sd = sd(modelo_ue$residuals)),
                size = 2, color = "grey30") +
  scale_color_manual(values = "grey50") +
  labs(x = "Resíduos",
       y = "Frequência") +
  theme_bw()
```



O teste de Shapiro-Francia indica que os resíduos não seguem a normalidade. Então é indicado aplicar Box-Cox para normalizar a variável Y e buscar por outros tipos de dependência entre X e Y:

```
lambda_ue <- powerTransform(dados_ue$cases_100k) #função powerTransform do pacote car#
lambda_ue
```

```
## Estimated transformation parameter
## dados_ue$cases_100k
##            0.256943
```

O parâmetro λ de Box-Cox igual à 0.25 indica a possibilidade de se obter um modelo utilizando esta técnica, devido à distância entre o parametro obtido e 1. O modelo com Box-Cox é construído conforme:

```
dados_ue$bc_cases <- (((dados_ue$cases_100k ^ lambda_ue$lambda) - 1) /
                        lambda_ue$lambda)
modelo_bc_ue <- lm(formula = bc_cases ~ tests_100k,
                   data = dados_ue)
summary(modelo_bc_ue)
```

```
##
## Call:
```

```
## lm(formula = bc_cases ~ tests_100k, data = dados_ue)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.534  -4.024   0.496   4.070  12.707
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.409e+01  4.863e-01  28.972   <2e-16 ***
## tests_100k  2.280e-05  1.439e-05   1.584    0.115
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.469 on 194 degrees of freedom
## Multiple R-squared:  0.01277,    Adjusted R-squared:  0.007683
## F-statistic:  2.51 on 1 and 194 DF,  p-value: 0.1148
```

Novamente, o resultados obtidos do modelo obtido indicam que este não tem significância estatística. Pois ambos os testes F e T resultaram em p-valores acima de um nível de significância de 5%.

Com esta análise, atesta-se que, para os dados europeus, os casos positivos identificados em cada país no período de tempo analisado não são afetados pela variação da testagem no mesmo período.

3 Discussão dos resultados

As análises realizadas neste estudo indicaram que, diferentemente dos dados europeus, os dados brasileiros mostraram uma forte relação entre o número de casos observados nos estados

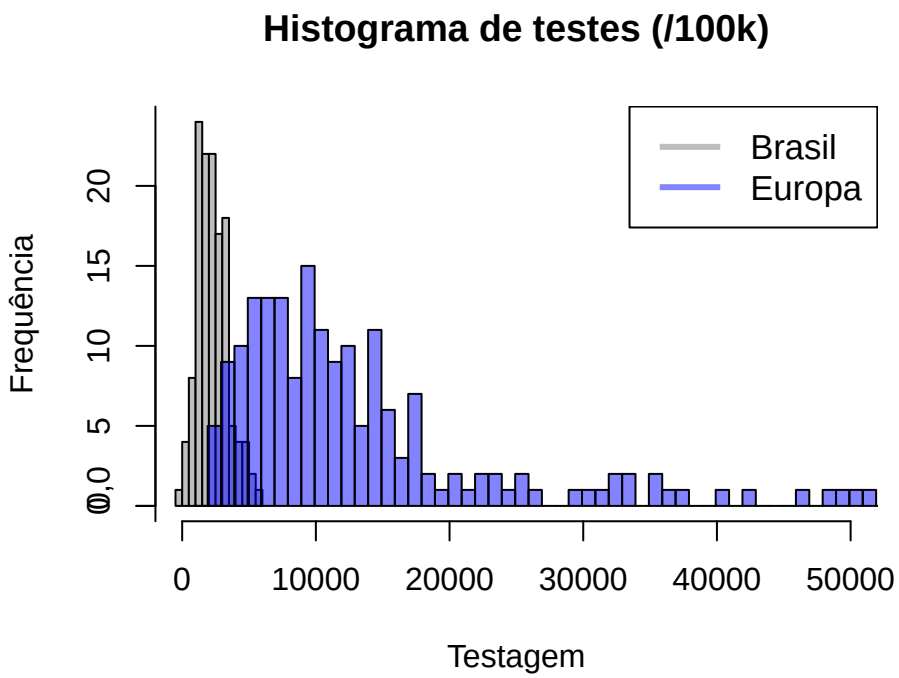
com a testagem realizada. De fato, um aumento na testagem resulta em um aumento médio no número de casos observados. Este construto validado pode levar a interpretação de que, no Brasil, o número de teste aplicados não é suficiente para avaliar a totalidade dos casos positivos, ou seja, existe uma alta sub-notificação. Pois, quando o número de testes aplicados já é suficiente, um aumento deste valor não deve levar a um aumento no número de casos positivos detecados, conforme é observado nos países europeus.

Esta interpretação pode ser amparada pela comparação entre a quantidade de testes (por 100 mil habitantes) aplicados no cenário brasileiro e no cenário europeu, conforme a seguinte imagem:

```
font_size=1.05
a=1000
hist(dados1$monthly_tests_100k,freq=TRUE,
     xlim=c(0,50000),right = FALSE,
     col="gray",main="Histograma de testes (/100k)",xlab="Testagem",
     ylab="Frequência",cex=font_size)
axis(2,at=pretty(dados1$monthly_tests_100k),
     labels=chartr(".", ",",
                   as.character(format(pretty(dados1$monthly_tests_100k),nsmall=1))))
hist(dados_ue$tests_100k,freq = TRUE,
     xlim=c(0,50000),right=FALSE,
     col=rgb(0, 0, 255, max = 255, alpha = 125, names = "blue50"),
     cex=font_size,add=T,
     breaks = seq(min(dados_ue$tests_100k)-a,max(dados_ue$tests_100k)+a,a))


legend("topright", c("Brasil","Europa"),
       col=c("gray",rgb(0, 0, 255, max = 255, alpha = 125, names = "blue50"),"black"),
```

```
       lwd=c(3,3,2),cex=font_size)
```



Observa-se então que a grande maioria dados observados na Europa apresentam uma massa de testagem bastante superior aos dados brasileiros. Este fato novamente sustenta o construto de que a testagem aplicada no Brasil é insuficiente, o que causa à dependência com caráter positivo entre a testagem e o número de casos observados nos dados brasileiros.