Atividade Monitorada IV

Gabriel Canfield & Patrick Zajdenwerg

Dados

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.0.3
```

```
dados <- read_excel("C:/Users/PatrickCamargo/Downloads/Atividade Monitorada IV - 2SEM2020.xl
View(dados)
```

Packages

```
library(ggplot2)
library(dplyr)
```

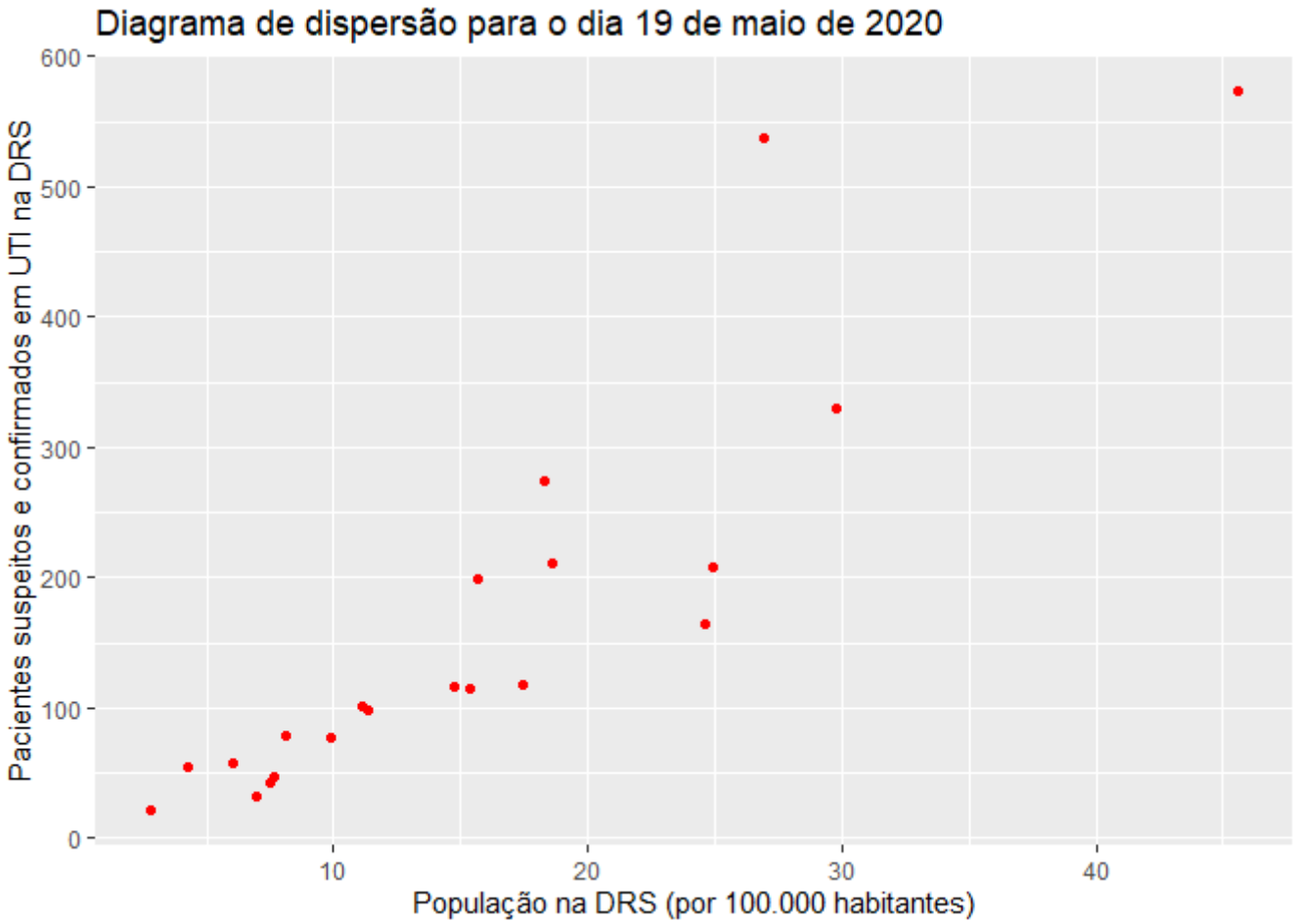
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Questão A

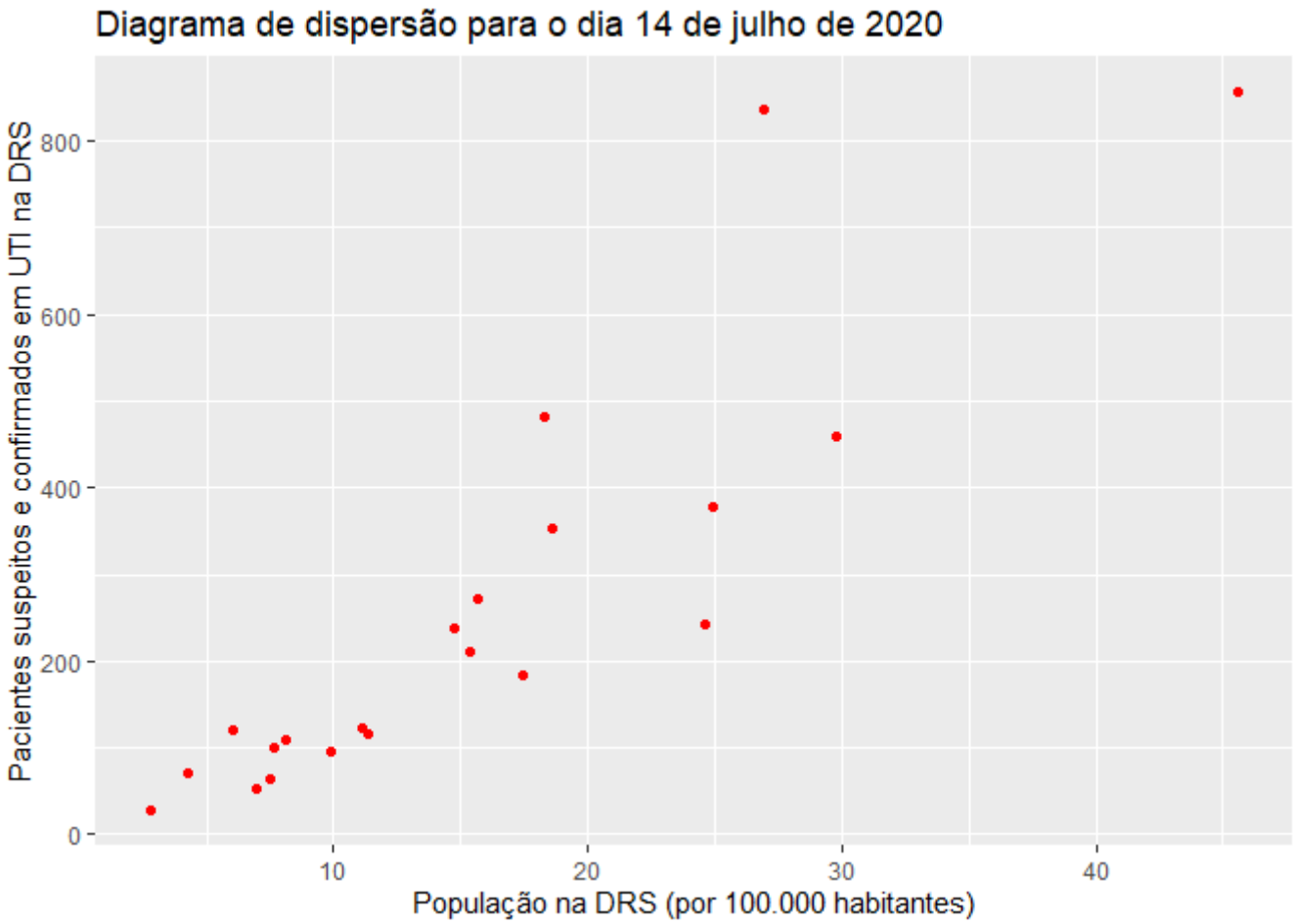
```
# seleção de dados para 19/05
dados_19_05 <- filter(dados, dia == "19 de maio de 2020")
x_pop_19_05 <- select(dados_19_05, pop)
y_pac_19_05 <- select(dados_19_05, total_uti)

# seleção de dados para 14/07
dados_14_07 <- filter(dados, dia == "14 de julho de 2020")
```

```
x_pop_14_07 <- select(dados_14_07, pop)
y_pac_14_07 <- select(dados_14_07, total_uti)
```

```
# diagrama de disperção 19/05
ggplot(dados_19_05) +
  geom_point(aes(x = pop, y = total_uti), color = "red") +
  xlab("População na DRS (por 100.000 habitantes)") +
  ylab("Pacientes suspeitos e confirmados em UTI na DRS") +
  ggtitle("Diagrama de dispersão para o dia 19 de maio de 2020")
```




```
# diagrama de disperção 14/07
ggplot(dados_14_07) +
  geom_point(aes(x = pop, y = total_uti), color = "red") +
  xlab("População na DRS (por 100.000 habitantes)") +
  ylab("Pacientes suspeitos e confirmados em UTI na DRS") +
  ggtitle("Diagrama de dispersão para o dia 14 de julho de 2020")
```



Pelos diagramas obtidos, é visualmente intuitiva a existência de um grau significativo de correlação entre as duas variáveis, em ambas as datas.

Questão B

```
# 19/05
attach(dados_19_05)
cor(total_uti, pop)
```

```
## [1] 0.8992129
```

```
modelo_19_05 <- lm(total_uti ~ pop)
summary(modelo_19_05)
```

```
##
## Call:
## lm(formula = total_uti ~ pop)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -120.15  -24.61   -9.35   21.36  221.16
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -43.77      27.68  -1.581     0.13
## pop            13.35       1.49   8.959    3e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.81 on 19 degrees of freedom
## Multiple R-squared:  0.8086, Adjusted R-squared:  0.7985
## F-statistic: 80.26 on 1 and 19 DF,  p-value: 2.999e-08
```

```
# 14/07
attach(dados_14_07)
```

```
## The following objects are masked from dados_19_05:
##
##     dia, nome_drs, pop, total_uti
```

```
cor(total_uti, pop)
```

```
## [1] 0.8893881
```

```
modelo_14_07 <- lm(total_uti ~ pop)
summary(modelo_14_07)
```

```
##
## Call:
## lm(formula = total_uti ~ pop)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -197.30  -45.59  -10.95   31.99  350.66
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -59.592     44.481   -1.34    0.196
## pop           20.306      2.395    8.48 6.98e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 110.6 on 19 degrees of freedom
## Multiple R-squared:  0.791,  Adjusted R-squared:   0.78
## F-statistic: 71.91 on 1 and 19 DF,  p-value: 6.976e-08
```

19/05/2020

Correlação: 0.8992129

R-quadrado: 0.8086

R-quadrado ajustado: 0.7985

14/07/2020

Correlação: 0.8893881

R-quadrado: 0.791

R-quadrado ajustado: 0.78

Os valores obtidos fundamentam a existência de correlação previstas pelas observações na questão A e, além disso, sugerem que a população é uma variável preditora para o número de pacientes suspeitos e confirmados em UTI na DRS.

Questão C

Equação de regressão para 19/05

Y = -43.77 + 13.35 * X

Teste de significância

H0: Beta-1 = 0

Ha: Beta-1 != 0

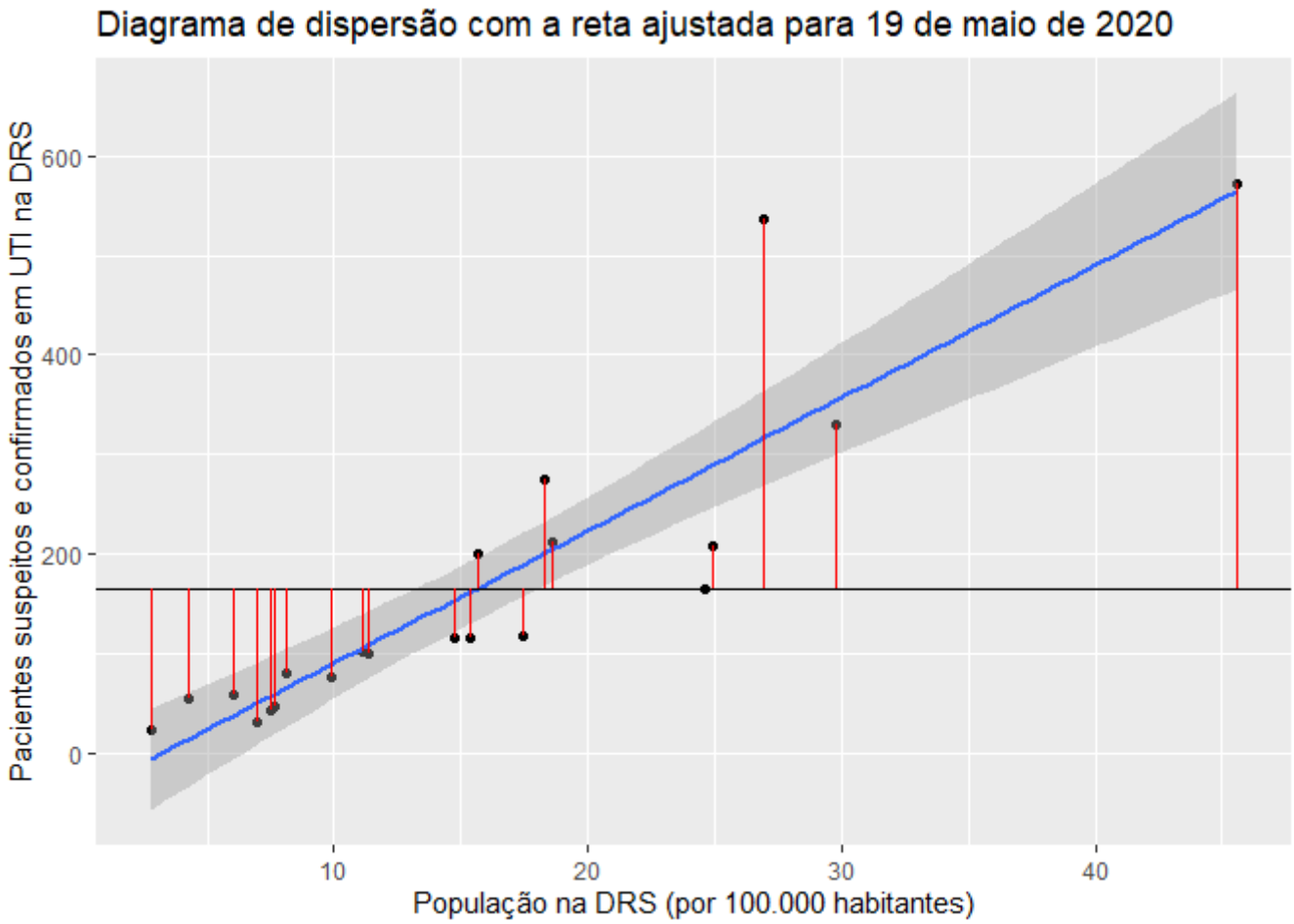
Como observado na análise da regressão, o F observado (80.26) é maior que o F crítico (2.99), rejeita-se H0. Portanto, ao nível de significância de 10%, os coeficintes são válidos.

```
attach(dados_19_05)
```

```
## The following objects are masked from dados_14_07:
##
##     dia, nome_drs, pop, total_uti

## The following objects are masked from dados_19_05 (pos = 4):
##
##     dia, nome_drs, pop, total_uti
```

```
ggplot(mapping = aes(pop, total_uti )) +
  geom_point() +
  geom_smooth(method = "lm") +
  geom_hline(yintercept = mean(total_uti)) +
  geom_segment(aes(x = pop  , y = total_uti, xend = pop , yend = mean(total_uti)), color="re
  xlab("População na DRS (por 100.000 habitantes)") +
  ylab("Pacientes suspeitos e confirmados em UTI na DRS") +
  ggtitle("Diagrama de dispersão com a reta ajustada para 19 de maio de 2020")
```

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



Equação de regressão para 14/07

Y = -59.592 + 20.306 * X

Teste de significância

H0: Beta-1 = 0

Ha: Beta-1 != 0

Como observado na análise da regressão, o F observado (71.91) é maior que o F crítico (2.99), rejeita-se H0. Portanto, ao nível de significância de

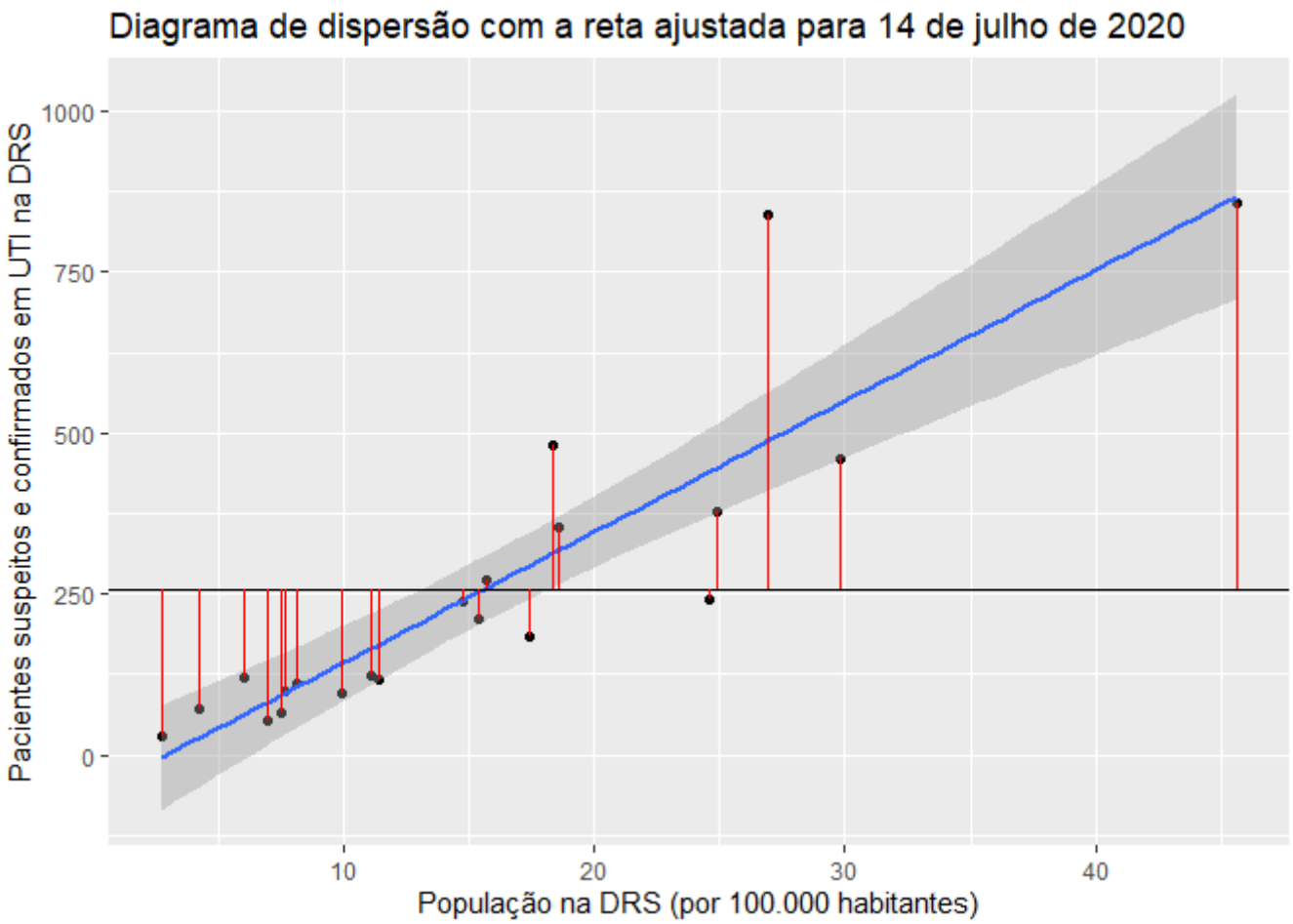
10%, os coeficintes são válidos.

```
attach(dados_14_07)
```

```
## The following objects are masked from dados_19_05 (pos = 3):
##
##     dia, nome_drs, pop, total_uti

## The following objects are masked from dados_14_07 (pos = 4):
##
##     dia, nome_drs, pop, total_uti

## The following objects are masked from dados_19_05 (pos = 5):
##
##     dia, nome_drs, pop, total_uti
```

```
ggplot(mapping = aes(pop, total_uti)) +
  geom_point() +
  geom_smooth(method = "lm") +
  geom_hline(yintercept = mean(total_uti)) +
  geom_segment(aes(x = pop  , y = total_uti, xend = pop , yend = mean(total_uti)), color="re
  xlab("População na DRS (por 100.000 habitantes)") +
  ylab("Pacientes suspeitos e confirmados em UTI na DRS") +
  ggtitle("Diagrama de dispersão com a reta ajustada para 14 de julho de 2020")
```

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




```
# Valores previstos pelo modelo em 19 de maio de 2020
previstos_19_05 <- predict(modelo_19_05, interval = "confidence", level = 0.90)
previstos_19_05
```

```
##           fit         lwr       upr
## 1  353.745541 308.9383130 398.55277
## 2   36.785186   0.9776357  72.59274
## 3  204.293526 177.2195511 231.36750
## 4  315.406852 276.3940312 354.41967
## 5  108.209784  80.0609279 136.35864
## 6   58.221349  25.1267089  91.31599
## 7   88.577411  58.7593580 118.39546
## 8  200.773518 173.8841738 227.66286
## 9   12.972937 -26.1418199  52.08769
## 10 188.678431 162.2996926 215.05717
## 11 565.248237 483.6664833 646.82999
## 12  49.183041  14.9793013  83.38678
## 13 104.361244  75.9172159 132.80527
## 14 161.754938 135.7851912 187.72468
## 15  56.648640  23.3649250  89.93235
## 16  -6.562251 -48.5716243  35.44712
## 17 153.468876 127.4168724 179.52088
## 18  64.334773  31.9585509  96.71099
## 19 165.869401 139.9039698 191.83483
## 20 284.859091 250.0251632 319.69302
## 21 288.579475 253.2627903 323.89616
```

```r
mean(previstos_19_05)
```

```
## [1] 164.5433
```

```r
# Valores previstos pelo modelo em 14 de julho de 2020
previstos_14_07 <- predict(modelo_14_07, interval = "confidence", level = 0.90)
previstos_14_07
```

```
##           fit       lwr       upr
## 1  545.086711 473.08267 617.09075
## 2   62.949867   5.40807 120.49166
## 3  317.751268 274.24410 361.25844
## 4  486.768710 424.07613 549.46129
## 5  171.595727 126.36125 216.83020
## 6   95.556986  42.37476 148.73921
## 7  141.732401  93.81557 189.64923
## 8  312.396891 269.18642 355.60737
## 9   26.728420 -36.12797  89.58481
## 10 293.998731 251.60879 336.38867
## 11 866.809101 735.70935 997.90885
## 12  81.808574  26.84406 136.77309
## 13 165.741610 120.03280 211.45042
## 14 253.044688 211.31198 294.77739
## 15  93.164696  39.67864 146.65076
## 16  -2.987075 -70.49504  64.52089
## 17 240.440538 198.57565 282.30543
## 18 104.856278  52.82854 156.88402
## 19 259.303307 217.57754 301.02908
## 20 440.301697 384.32448 496.27891
## 21 445.960873 389.20788 502.71386
```

```r
mean(previstos_14_07)
```

```
## [1] 257.2862
```

O pior cenário é em 14/07, uma vez que os valores previstos de ocupação de UTI são maiores.

Questão D

```r
# definição do data frame para 1 milhão de habitantes
new_1mm = data.frame(pop = 10)
```

```r
# 19/05
predict(modelo_19_05, new_1mm, interval = "confidence", level = 0.90)
```

```
##       fit      lwr     upr
## 1 89.7208 60.01064 119.431
```

```r
# 14/07
predict(modelo_14_07, new_1mm, interval = "confidence", level = 0.90)
```

```
##        fit      lwr      upr
## 1 143.4716 95.72819 191.2151
```

Questão E

```r
# definição do data frame para 500 mil habitantes
new_500k = data.frame(pop = 5)
```

```r
# 19/05
predict(modelo_19_05, new_500k, interval = "confidence", level = 0.90)
```

```
##        fit       lwr      upr
## 1 22.97308 -14.71989 60.66605
```

```r
# 14/07
predict(modelo_14_07, new_500k, interval = "confidence", level = 0.90)
```

```
##       fit       lwr      upr
## 1 41.9399 -18.63171 102.5115
```

Questão F

Análise 19 de maio de 2020

```r
# teste de sigmificância
anova(modelo_19_05)
```

```
## Analysis of Variance Table
##
## Response: total_uti
##           Df Sum Sq Mean Sq F value    Pr(>F)
## pop        1 380024  380024   80.26 2.999e-08 ***
## Residuals 19  89963    4735
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

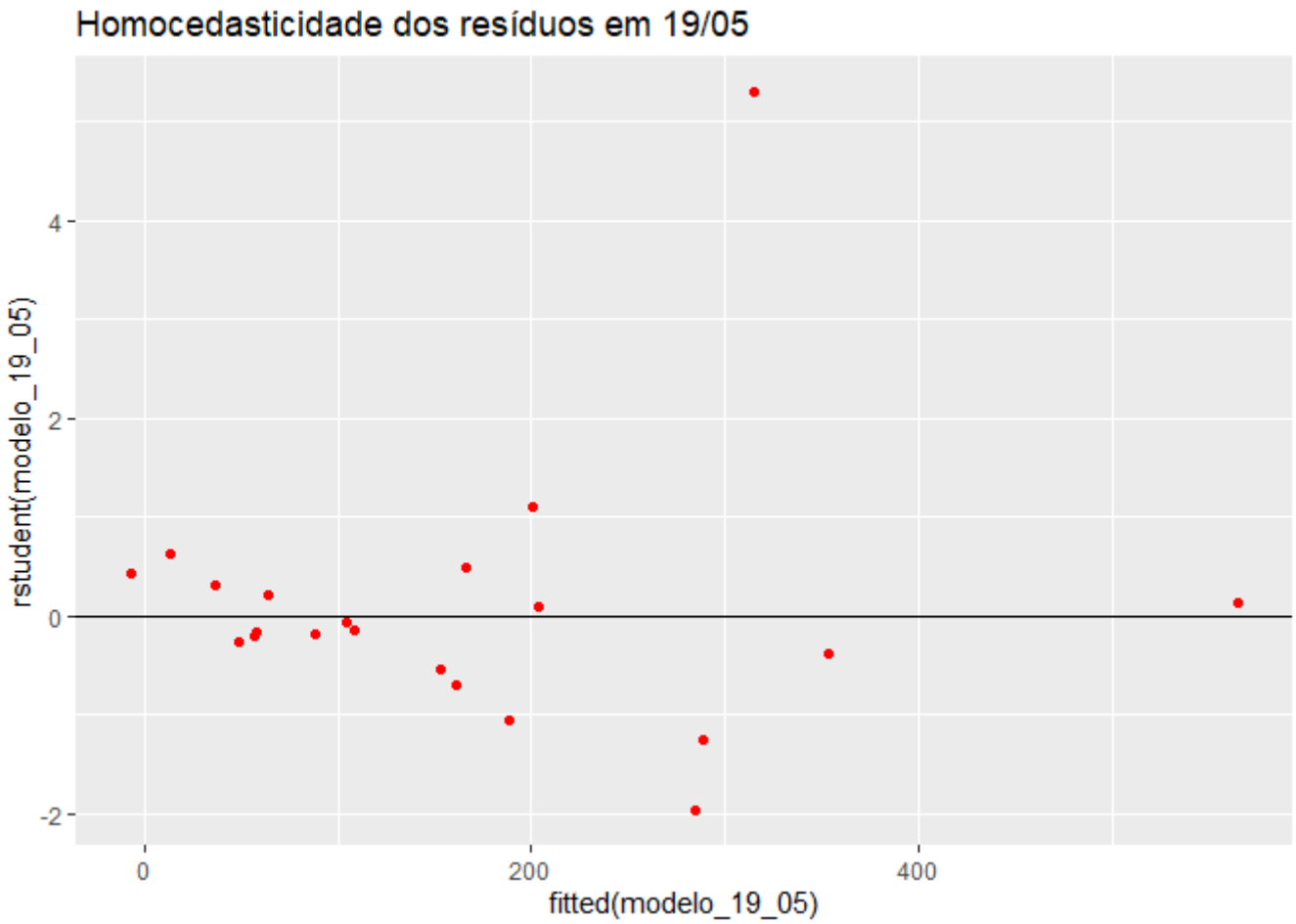
O p-value é quase nulo e tem um valor menor que 0.10, ou seja, a relação entre as duas variáveis é significativa.

```
# linearidade
cor(dados_19_05$total_uti, dados_19_05$pop)
```

```
## [1] 0.8992129
```

Existe uma correlação forte entre as variáveis.

```
# homocedasticidade dos resíduos
ggplot(modelo_19_05) +
  geom_point(aes(fitted(modelo_19_05), rstudent(modelo_19_05)), color = "red") +
  ggtitle("Homocedasticidade dos resíduos em 19/05") +
  geom_hline(yintercept = 0)
```



Podemos observar que os resíduos se distribuem de maneira aleatória ao redor dos valores previstos, fazendo com que o modelo passe nesse critério de avaliação.

```
shapiro.test(modelo_19_05$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  modelo_19_05$residuals
## W = 0.86331, p-value = 0.007266
```

Como o p-value = 0.007266 < 0.1, concluímos que para um nível de significância de 10%, nossos dados possuem uma distribuição normal.

Análise 14 de julho de 2020

```
# teste de sigmificância
anova(modelo_14_07)
```

```
## Analysis of Variance Table
##
## Response: total_uti
##           Df Sum Sq Mean Sq F value    Pr(>F)
## pop        1 879310  879310  71.914 6.976e-08 ***
## Residuals 19 232318   12227
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

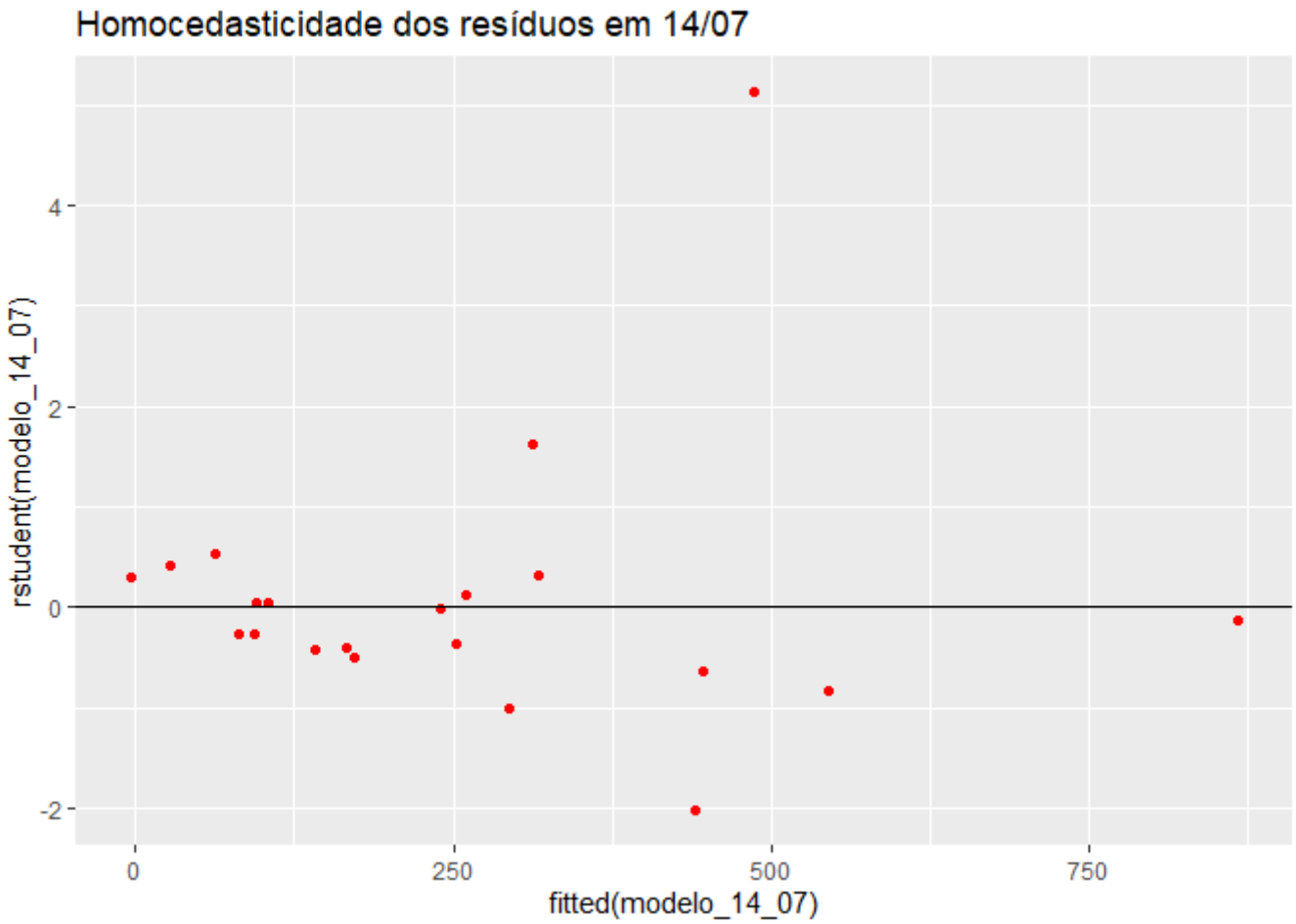
O p-value é quase nulo e tem um valor menor que 0.10, ou seja, a relação entre as duas variáveis é significativa.

```
# linearidade
cor(dados_14_07$total_uti, dados_14_07$pop)
```

```
## [1] 0.8893881
```

Existe uma correlação forte entre as variáveis.

```
# homocedasticidade dos resíduos
ggplot(modelo_14_07) +
  geom_point(aes(fitted(modelo_14_07), rstudent(modelo_14_07)), color = "red") +
  ggtitle("Homocedasticidade dos resíduos em 14/07") +
  geom_hline(yintercept = 0)
```



Podemos observar que os resíduos se distribuem de maneira aleatória ao redor dos valores previstos, fazendo com que o modelo passe nesse critério de avaliação.

```
shapiro.test(modelo_14_07$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  modelo_14_07$residuals
## W = 0.84574, p-value = 0.00359
```

Como o p-value = 0.00359 < 0.1, concluímos que para um nível de significância de 10%, nossos dados possuem uma distribuição normal.

Para ambas as datas, o modelo apresenta um valor negativo para a soma de pacientes suspeitos e pacientes confirmados em UTI na DRS 12 Registro, o que é impossível de ocorrer uma vez que não há como os números de internações serem negativos. Portanto, há a presença de pontos influentes.