

Análise estatística descritiva com uso de R

Análise bidimensional: Correlação linear e chi-quadrado

Marília Melo Favalesso



Hola!

Marília Melo Favalessos



✉ mariliabio@ufpr@gmail.com

🌐 www.mmfava.com

⌚ github.com/mmfava

Educação

🎓 Estudante de doutorado em '*'Ecología, Genética y Evolución'*', 2018 - atual, *Universidad de Buenos Aires*, Argentina.

🎓 Mestre em Conservação e Manejo de Recursos Naturais, 2016 - 2018, Universidade estadual do Oeste do Paraná, Brasil.

🎓 Bacharela em Ciências Biológicas, 2010 - 2014, Universidade Federal do Paraná, Brasil.

🎓 Técnica em Meio-ambiente, 2005 - 2009, CEEP Pedro Boaretto Neto, Brasil.

Lembre-se!

Pergunte, pergunte & perguntas!   

Aula de revisão de conteúdos e exercícios! 

De preferência, mantenha a câmera ligada 

Respeitar os colegas e a profa. 

Tentar se divertir durante a aula. 

Conteúdo programático da mini-aula

- O que é correlação
 - Correlação ≠ Causalidade
- Correlação linear
 - Coeficiente de correlação produto-momento (r)
 - Exercício
- χ^2 de associação
 - Resíduos ajustados
 - Exercício

O que é correlação?

O que é correlação?

Em estatística, a **correlação** mede a **direção** e a **intensidade** de uma relação entre duas variáveis.

- Por exemplo:

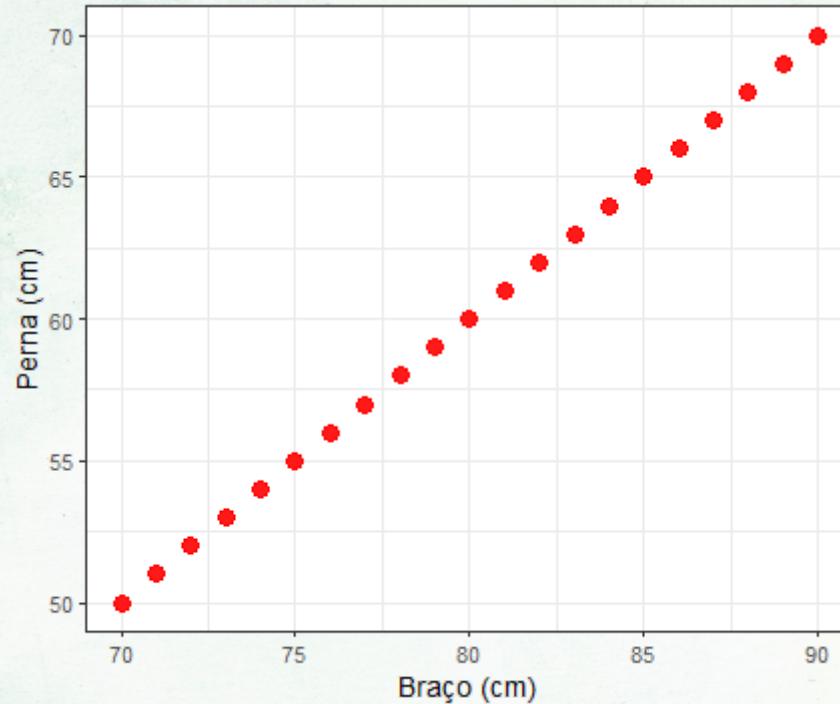


Quando se pode demonstrar que existe associação entre duas variáveis, isto é, quando se constata que elas variam juntas, diz-se que as variáveis estão **correlacionadas**.

O que é correlação?

Correlação ≠ Causalidade

↑ ↘ ~ ↑ 🧑 ?

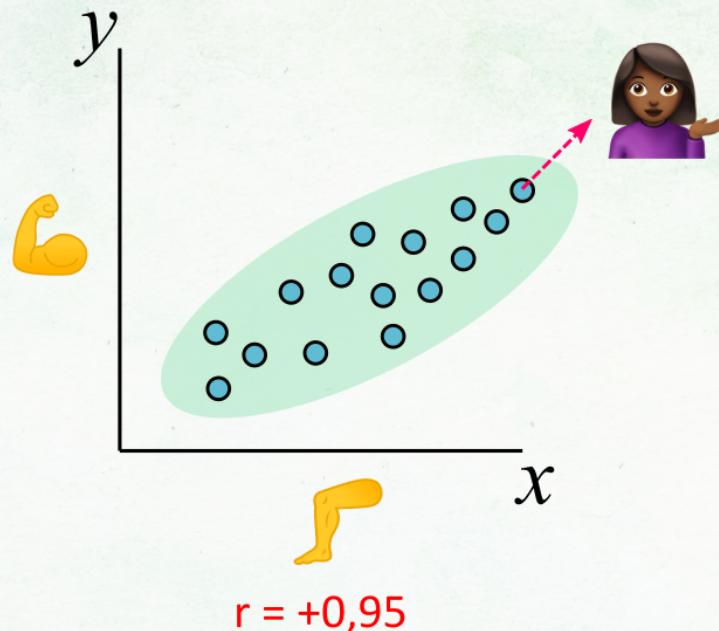


Correlação linear

Correlação linear

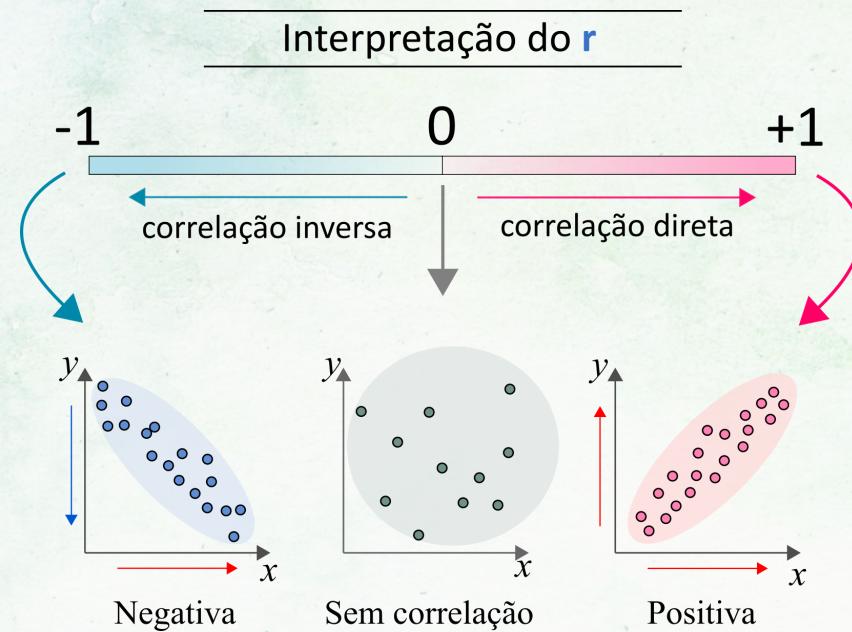
Coeficiente de correlação produto-momento (r)

- O r (ρ) é uma medida da força e direção da correlação entre duas variáveis quantitativas e também é conhecido como **coeficiente de correlação de Pearson**.



Correlação linear

Direção da correlação



A partir dos diagramas de dispersão, já temos uma indicação se a relação entre as variáveis é **positiva** (quando as duas crescem juntas), **negativa** (quando uma aumenta e a outra diminui) ou se **não há correlação**.

Correlação linear

Intensidade da correlação

Interpretação do $|r|$

0.0	Nula
0.0 -- 0.3	Fraca
0.3 - 0.6	Regular
0.6 - 0.9	Forte
0.9 - 1.0	Muito forte
1.0	Plena ou perfeita

Correlação linear

Exercício

Existe associação entre o PIB total (R\$) dos municípios brasileiros e o tamanho de suas populações?

```
## Pacotes ----
library(tidyverse) # manipulação dos dados
library(janitor) # limpeza dos dados

## Importar os dados ----
# • Para importar os dados em .csv, utilizamos a função "read.csv".
dados_socio_economicos <-
  read_csv("dados_socio_economicos.csv") %>% # importação da tabela
  clean_names # examinar e limpar os nomes das colunas

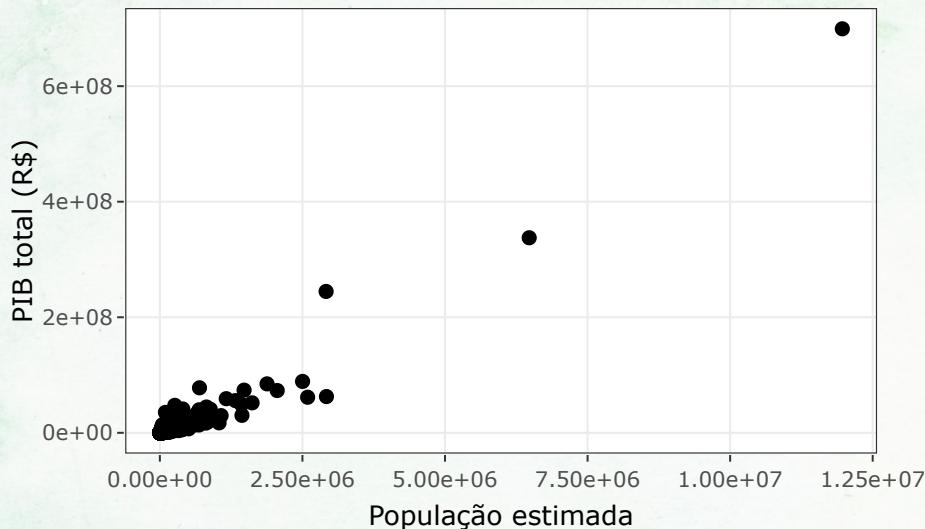
dados_socio_economicos %>% names # Consultar as variáveis contidas na tabela

## [1] "uf" "cod_uf" "cod_municipio" "nome_do_municipio" "populacao_estimada" "pib_per_capita" "pib_total_em_r_mi"
## [2] "atividade_economica" "ranking_pib_per_capita" "ranking_pib_total"
```

Correlação linear

Exercício

```
## Plotar os dados em um gráfico de dispersão ---  
dados_socio_economicos %>% # tabela de dados  
  ggplot(aes(x = populacao_estimada, y = pib_total_em_r_mi)) + # var linha e coluna x  
  geom_point() + # transparência e tamanho dos pontos  
  xlab("População estimada") + # nome do eixo X  
  ylab("PIB total (R$)") # nome do eixo Y
```



Correlação linear

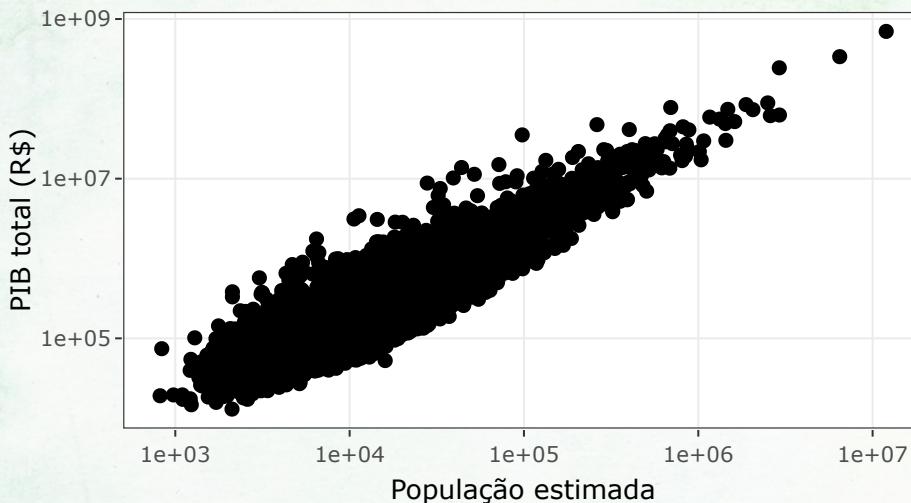
Exercício

```
## Plotar os dados com transformação nos eixos ---  
dados_socio_economicos %>%  
  ggplot(aes(x = populacao_estimada, y = pib_total_em_r_mi)) +  
    geom_point() +  
    scale_x_log10() +  
    scale_y_log10() +  
    xlab("População estimada") + ylab("PIB total (R$)")
```

Correlação linear

Exercício

```
## Plotar os dados com transformação nos eixos ---  
dados_socio_economicos %>%  
  ggplot(aes(x = populacao_estimada, y = pib_total_em_r_mi)) +  
  geom_point() +  
  scale_x_log10() + # transformação log dados do eixo x  
  scale_y_log10() + # transformação log dados do eixo y  
  xlab("População estimada") + ylab("PIB total (R$)")
```



Correlação linear

Exercício

- Para finalizar, precisamos calcular o **r** para os nossos dados:

```
cor(  
  dados_socio_economicos$populacao_estimada,  
  dados_socio_economicos$pib_total_em_r_mi  
)  
## [1] 0.9621635
```

Existe uma correlação _____ e _____ entre a população estimada e o PIB total ($r =$ _____).

Correlação linear

Exercício

- Para finalizar, precisamos calcular o **r** para os nossos dados:

```
cor(  
  dados_socio_economicos$populacao_estimada,  
  dados_socio_economicos$pib_total_em_r_mi  
)  
## [1] 0.9621635
```

Existe uma correlação **positiva** e **muito forte** entre a população estimada e o PIB total ($r = 0,96$).

χ^2 de associação

Resíduos ajustados

χ^2 de associação

Resíduos ajustados

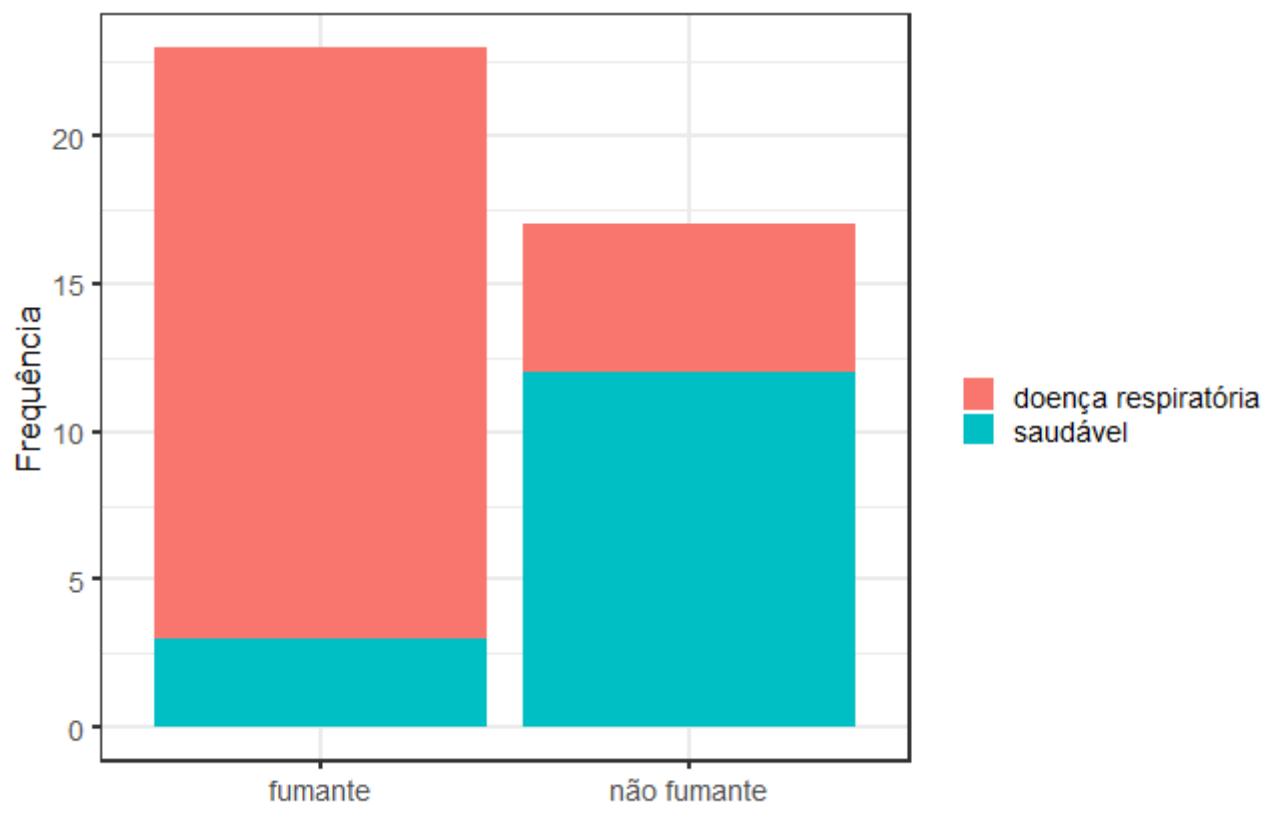
- Os resíduos do χ^2 (qui-quadrado) de associação ou de independência são utilizados para avaliar a correlação entre **variáveis categóricas**, assim como o coeficiente **r** é usado para o mesmo fim com variáveis quantitativas.
- **Exemplo:**
 - A presença de doenças respiratórias está associado ao tabagismo? (amostra: n = 40).



Doença respiratória	Fumante	
	Sim	Não
Sim	20	5
Não	3	12

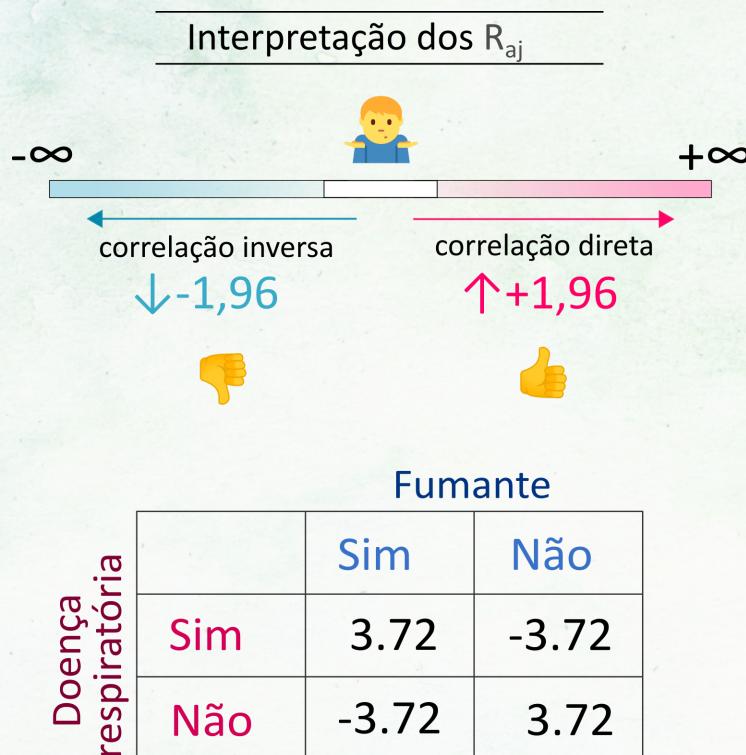
χ^2 de associação

Gráfico de barras empilhadas



χ^2 de associação

Resíduos ajustados



No R usamos a função `chisq.test(X, Y)$stdres` para obter os R_{aj} .

χ^2 de associação

Exercício

- Existe associação entre raça e identificação partidária para os entrevistados pelo *US General Social Survey*?

```
## Pacotes ----
library(tidyverse) # manipulação de dados + gráficos
library(janitor) # limpeza nos dados
library(dados) # onde está a tabela de dados
library(viridis) # para usar a paleta de cores viridis
library(corrplot) # usaremos para plotar os resíduos

## Chamar os dados da tabela 'questionario'{dados} ----
questionario <- questionario %>% clean_names

questionario %>% names # variáveis da tabela

## [1] "ano"          "estado_civil" "idade"        "raca"         "renda"        "partido"      "religiao"
## [8] "denominacao" "horas_tv"
```

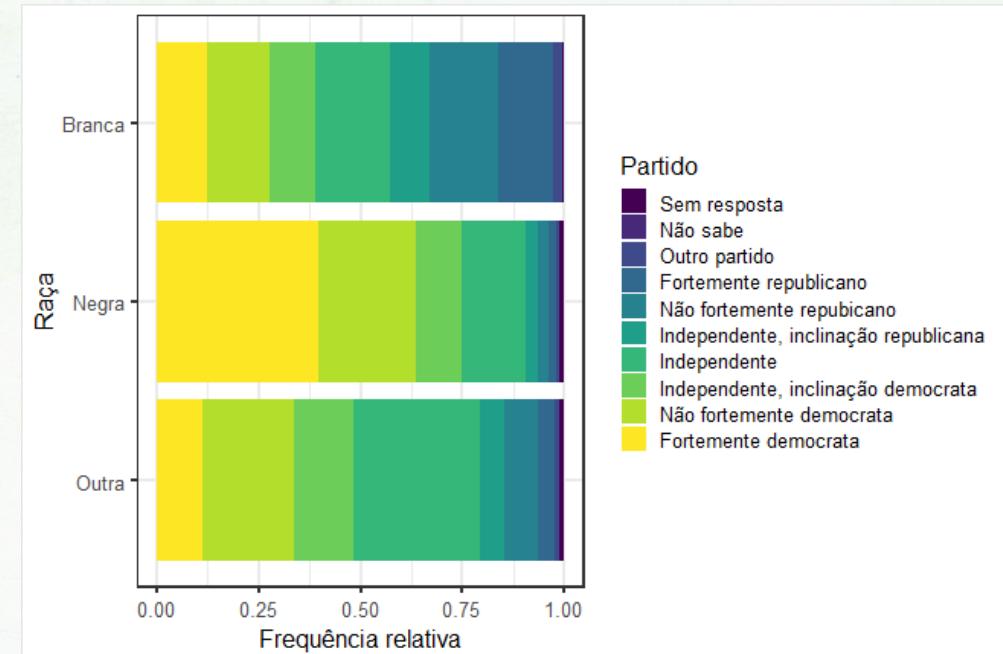
χ^2 de associação

Exercício

Gráfico de barras empilhadas

```
## Para criar o gráfico usamos o pacote 'ggplot'

questionario %>% # incluimos nossos dados
  ggplot(aes(y = raca, # raça vai no eixo "y"
             fill = partido)) + # as barras empilhadas
  geom_bar( # usamos a função "geom_bar" para criar o
    position = "fill" # posição das barras no gráfico
  ) +
  scale_fill_viridis( # usamos a paleta de cores 'viridis'
    discrete = TRUE, # ... para dados discretos
    name = "Partido") + # título da legenda
  xlab("Frequência relativa") + # nome do eixo x
  ylab("Raça") # nome do eixo y
```

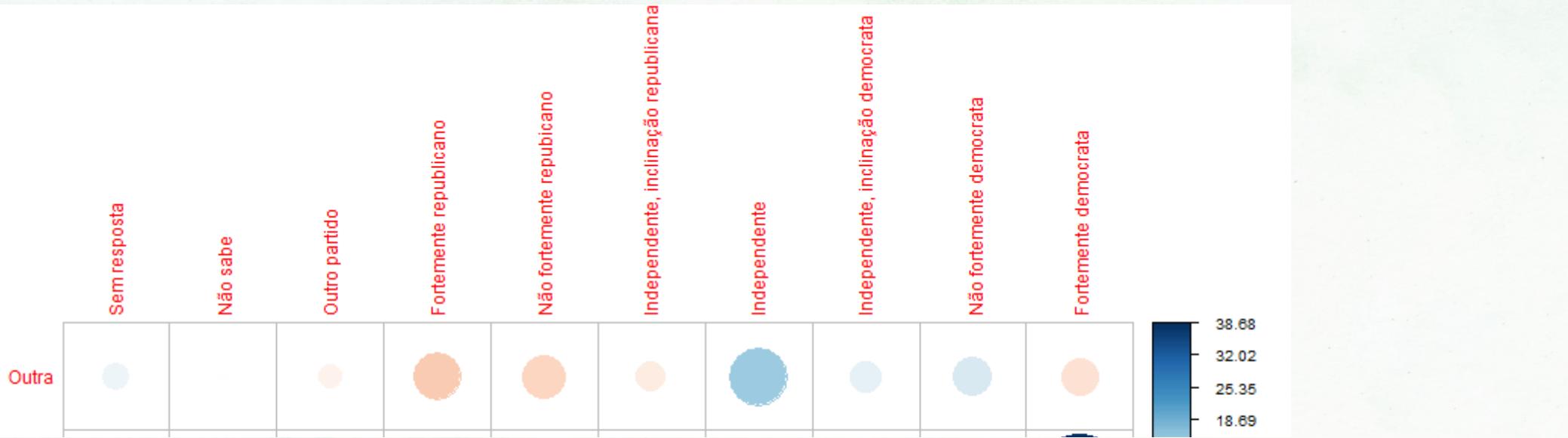


χ^2 de associação

Exercício

Análise dos resíduos

```
chisq.test(questionario$raca,  
          questionario$partido)$stdres %>%  
corrplot(is.cor = FALSE)
```



Dúvidas?



<https://www.menti.com/hs6etjj2u9> (cod: 7513-1177)

FIM!



THANK YOU