

Gráficos Contemporâneos

UnB - Universidade de Brasília

30 de outubro de 2019

Correlograma

O que é uma matriz de correlação?

É uma tabela que mostra o nível de correlação (Pearson, Spearman, Kendall) entre variáveis quantitativas/ordinais. Veja um exemplo a seguir:

Tabela 1: Correlação linear de Pearson entre variáveis do dataset Auto

	mpg	displacement	horsepower
mpg	1.0000000	-0.8051269	-0.7784268
displacement	-0.8051269	1.0000000	0.8972570
horsepower	-0.7784268	0.8972570	1.0000000

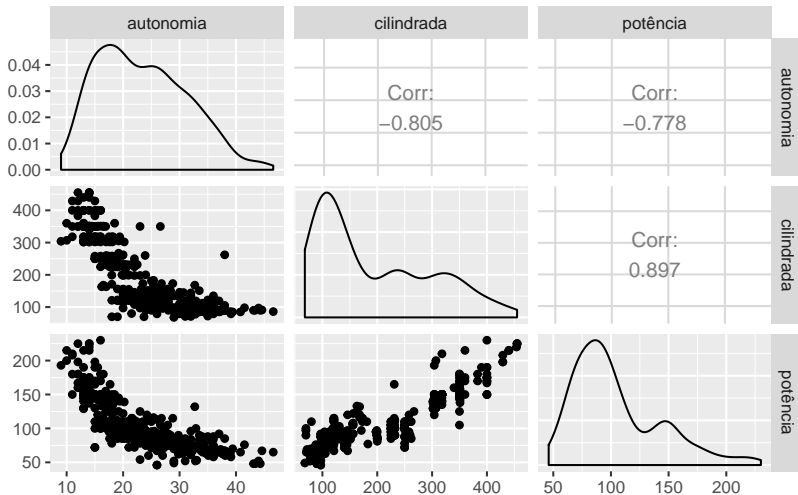
- ▶ Simétrica em torno da diagonal principal
- ▶ Diagonal principal irrelevante

O que é o correlograma?

O correlograma é uma representação gráfica para matrizes de correlação. Permite explorar alguns outros aspectos dos dados, incorporando múltiplos gráficos e cores, além das medidas de correlação em si.

Correlograma com o pacote GGally

correlograma no dataset Auto



Correlograma com o pacote GGally

```
library(GGally)

ggpairs(data = auto_dataset,
        columns = c("mpg", "displacement", "horsepower"),
        columnLabels = c("autonomia",
                          "cilindrada",
                          "potência"),
        title = "correlograma no dataset Auto")
```

data: data frame contendo os dados;

columns: indica quais variáveis serão selecionadas para o gráfico;

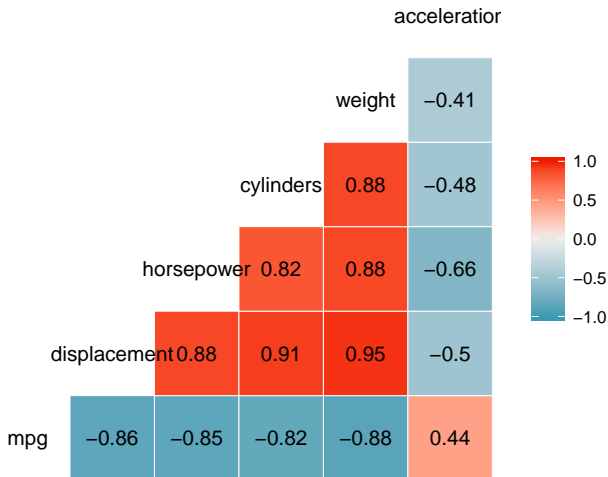
columnLabels: rótulos para as respectivas variáveis;

title: título do gráfico.

Obs: Com este pacote é ainda possível escolher os tipos de gráficos mostrados no correlograma (densidade/histograma/barras) de acordo com o tipo de variável (contínua/discreta), além das medidas de correlação, tanto para a parte acima quanto abaixo da diagonal principal.

Em seguida veremos uma outra forma de correlograma, feito pelo mesmo pacote. Este permite explorar um número muito maior de variáveis, resumindo a matriz de correlação em uma grade multicolorida.

Correlograma com o pacote GGally



Como implementar?

```
library(GGally); library(dplyr)

auto_dataset %>%
  select(mpg, displacement, horsepower,
         cylinders, weight, acceleration) %>%
  ggcorr(data = ., method = c("everything", "spearman"),
         label=TRUE, label_round=2,
         palette = "grey50")
```

data: data frame contendo os dados;

method: vetor contendo o tipo de tratamento aos dados faltantes e a medida de correlação que será usada;

label: ativa/desativa o aparecimento do valor das correlações;

label_round: número de casas decimais no label;

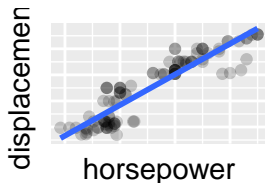
palette: esquema de cores.

Precauções

- ▶ Variáveis com nível de mensuração pelo menos ordinal;
- ▶ Escolha adequada das medidas de correlação;
- ▶ Tratar dados faltantes (NAs);
- ▶ Rótulos curtos para as variáveis.

Usos do correlograma / Aplicação

- ▶ Encontrar relações entre variáveis a serem exploradas



- ▶ Estabelecer hipóteses e inferir sobre correlações (aplicar teste de correlação)

$$H_0: \text{corr}(\text{disp}, \text{mpg}) = 0$$

$$H_1: \text{corr}(\text{disp}, \text{mpg}) < 0$$

- ▶ Identificar uma *possível* relação causal (aplicar regressão linear múltipla)

Ex: veículos com mais cilindros consomem mais combustível?