SME0820 - Modelos de Regressão e Aprendizado Supervisionado I - Trabalho I

Brenda da Silva Muniz 11811603 Francisco Rosa Dias de Miranda 4402962
Heitor Carvalho Pinheiro 11833351- Mônica Amaral Novelli 11810453

Setembro 2021

Neste trabalho, nosso objetivo é ajustar um modelo de regressão linear simples ao conjunto de dados fornecido, utilizando linguagem R. Para esta tarefa, descreveremos cada etapa de nosso *pipeline*.

Primeiramente, vamos carregar os módulos utilizados nesta análise. Caso não possua algum dos pacotes, utilize o comando `install_packages("Nome_do_pacote")`.

```
library(tidyverse)
library(ggpubr)
library(corrplot)
library(DataExplorer)
library(GGally)
library(knitr)
library(data.table)
```

Com os pacotes carregados em nosso ambiente, lemos o arquivo `.csv` disponibilizado colocando-o na mesma pasta de nosso projeto. Vamos inspecionar o que foi carregado com auxílio do comando `head()`, que exibe as 5 primeiras observações.

```
dados <- read_csv("data-table-B3.csv", locale = locale(decimal_mark = ","))

head(dados)
```

```
## # A tibble: 6 x 12
##       y    x1    x2    x3    x4    x5    x6    x7    x8    x9   x10   x11
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  18.9   350   165   260  8     2.56     4     3  200.  69.9  3910     1
## 2  17     350   170   275  8.5   2.56     4     3  200.  72.9  3860     1
## 3  20     250   105   185  8.25  2.73     1     3  197.  72.2  3510     1
## 4  18.2   351   143   255  8     3        2     3  200.  74    3890     1
## 5  20.1   225    95   170  8.4   2.76     1     3  194.  71.8  3365     0
## 6  11.2   440   215   330  8.2   2.88     4     3  184.  69    4215     1
```

Parte a):

- Descrição do banco de dados
- Definição das variáveis
- Análise exploratória inicial

```
ggcorr(dados, geom = "circle")
```

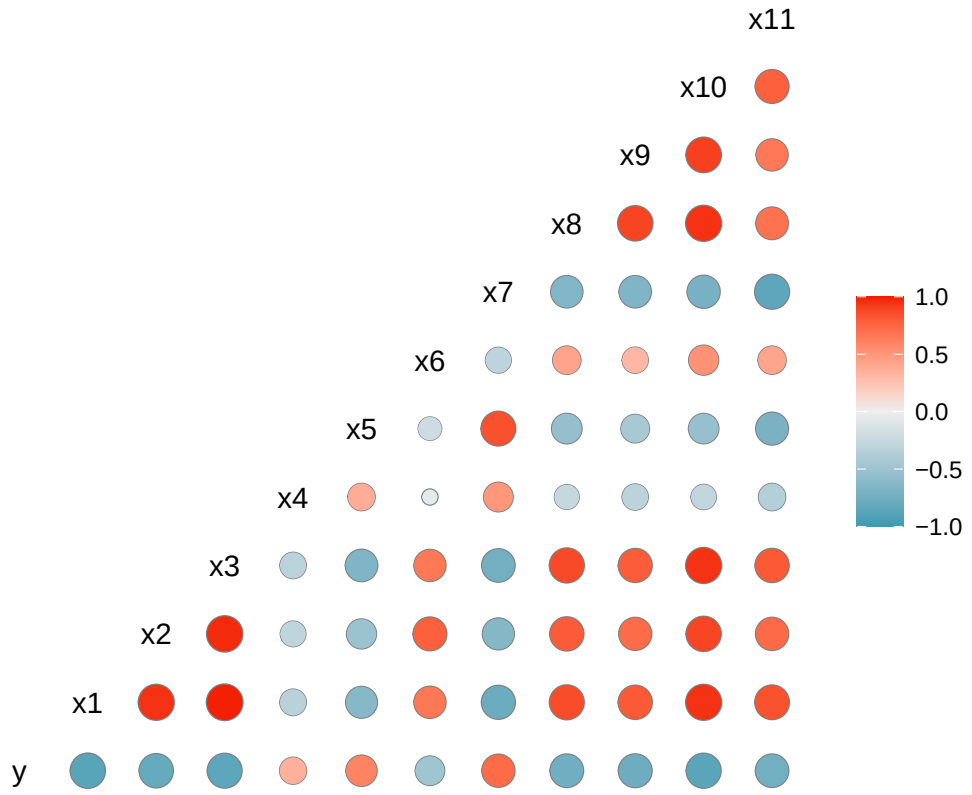
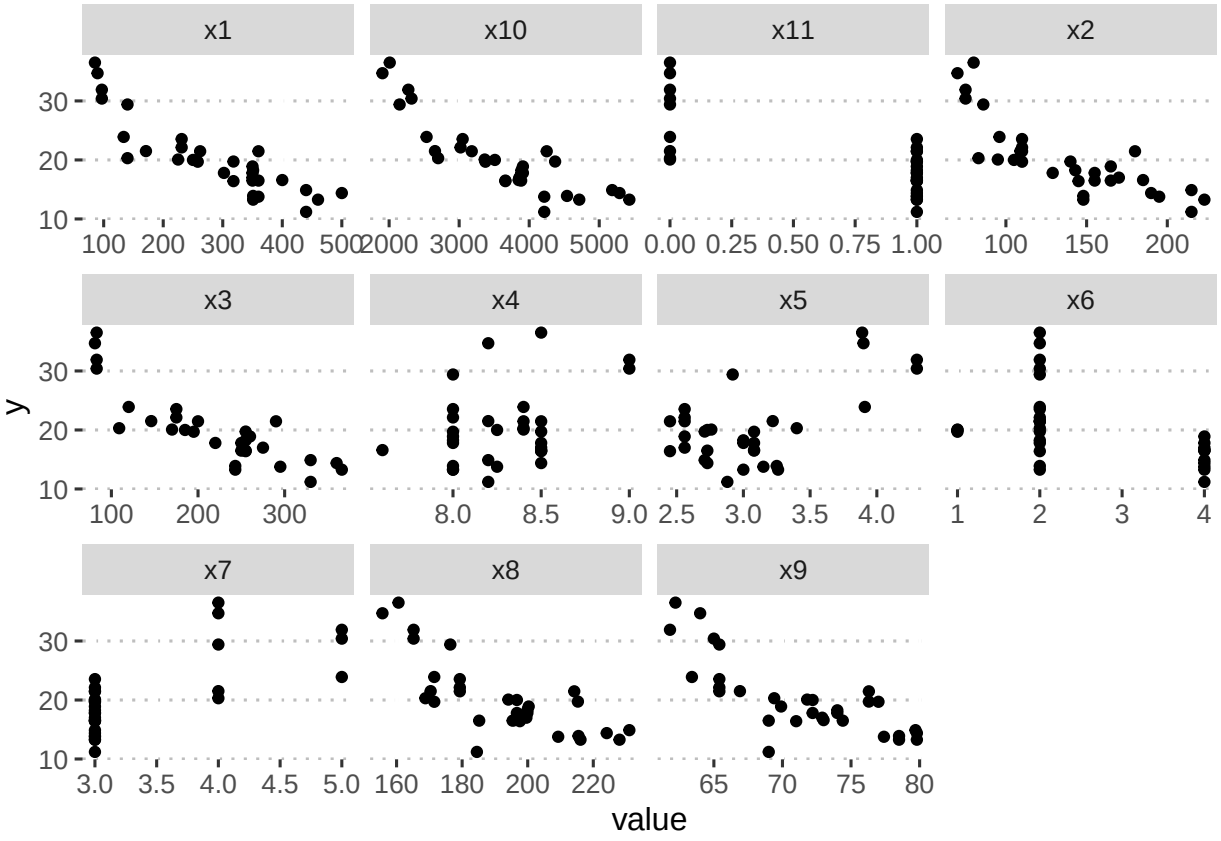


Figure 1: Correlograma entre as variáveis

- Graficos de dispersão Y versus X_i, $i = 1, ..., 11$.

```
dados %>%
  pivot_longer(cols = !"y") %>%
  ggplot(aes(y = y)) +
    geom_point(aes(x = value)) +
    facet_wrap(~name, scales = "free_x") + theme_pubclean()
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



Interpretação de cada gráfico

Parte b):

Consultar e descrever brevemente os conceitos Data splitting, cross validation, overfitting, underfitting, missing data, encoding data.

Parte c):

1. Calcular S_{XX},S_{YY} e S_{XY}
2. Ajustar um modelo de regressão linear simples, apresentar a estimativa de β_0, β_1 e σ^2 e fazer um gráfico com a reta ajustada

```
fit <- lm(y ~ x10, data = dados)

anova(fit)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x10        1 921.53  921.53  87.482 2.121e-10 ***
## Residuals 30 316.02   10.53
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3. Calcule o valor dos $\hat{Y}$ e o valor dos resíduos para seu modelo, resumo e histograma dos resíduos, e faça uma análise da distribuição destes.
4. testes de hipotese para β_0 e β_1
5. intervalos de confiança
6. intervalos de predição
7. tabela anova