LABORATÓRIO 2: Regressão Linear Simples

Fernando Bispo, Jeff Caponero

Sumário

Segunda parte

Apresentação

Com base nos dados sobre a eleição presidencial de 2000 nos Estados Unidos, referentes ao número de votos de cada um dos candidatos por condado no estado da Flórida. Deseja-se investigar a relação entre o número de votos que Bush recebeu em relação ao número de votos recebidos por Buchanan, bem como, trazer um pouco de luz sobre o debate referente aos votos recebidos por Buchanan que poderiam ter sido de Al Gore, se o primeiro não estivesse no pleito. Como Bush e Gore foram os candidatos principais daquela eleição, é de interesse avaliar a relação entre os votos recebidos por Bush e Buchanan na Flórida, que é um Estado importante na corrida presidencial dos EUA. Para isto, ajuste um modelo de regressão linear no qual o número de votos de Bush é usado para predizer o número de votos de Buchanan. Os dados estão disponíveis no arquivo "florida.csv".

Objetivos

Com os dados deste estudo:

(a) Discutir a relação entre os votos recebidos por Bush e por Buchanan através do uso de diagramas de dispersão.

(b) Avaliar a relevância do argumento que os votos de Buchanan poderiam ser direcionados a Al Gore, caso Buchanan não tivesse participado do pleito.

(c) Tratar dados atipicos.

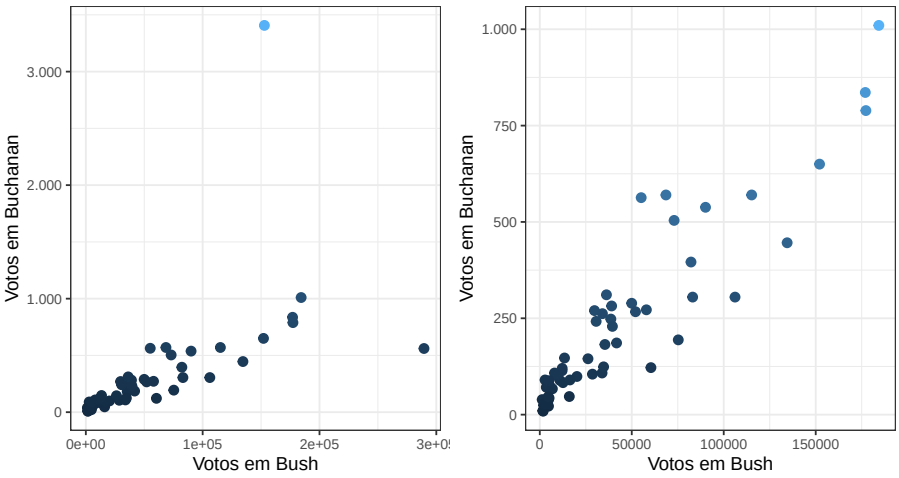
(d) Obter uma reta ajustada aos dados tratados e predizer a votação de Buchanan caso Bush receba 152.846 votos em determinado condado.

(e) Criar um programa baseado em estimativas de mínimos quadrados, prever a votação de Buchanan, sob as mesmas condições e compará-las.

Análise dos dados

O gráfico de dispersão sugere certa correlação positiva entre os votos de Bush e Buchanan, muito embora se observe que no condado de Palm Beach a votação de Buchanan (3.407 votos) represente um dado inesperado, bem como, em menor medida, a votação de Bush em Dade (289.456 votos). Retirados estes valores discrepantes, esta provável correlação parece ainda mais certa.

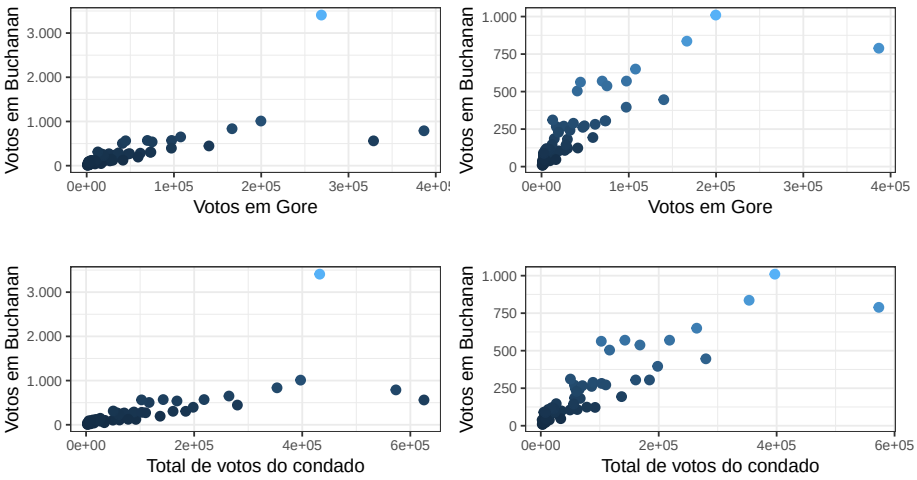
Figura 1: Relação entre votos recebidos por Bush e por Bachanan nos condados da Fló



Fonte: 2000 OFFICIAL PRESIDENTIAL GENERAL ELECTION RESULTS – USA

Embora a Figura 1 sugira certa correlação positiva entre os votos de Bush e Buchanan, a análise destes mesmos votos em relação a votação recebida por Al Gore e os votos totais dos condados parece retratar outra realidade (ver Figura 2), qual seja, que a correlação positiva é mais propriamente deviada a um fator externo, e não avaliado inicialmente, que corresponde ao aumento de eleitores nos condados. Desta forma, não parece ser plausível afirmar que a votação dada ao candidato Buchanan seria direcionada a qualquer dos candidatos caso este não participasse do pleito.

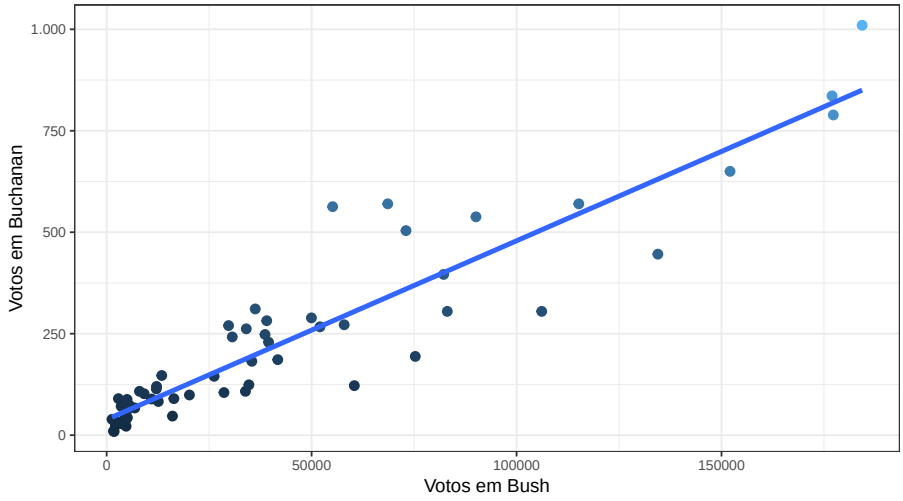
Figura 2: Relação entre votos recebidos por Gore e por Bachanan e com o total de voto



Fonte: 2000 OFFICIAL PRESIDENTIAL GENERAL ELECTION RESULTS – USA

Assumindo que haja uma correlação entre os votos dos dois candidátos é possível realizar uma regressão linear a partir dos dados dos votos dos condados. A Figura 3 mostra a reta de regressão linear obtida. Com base nesta regressão é possível inferir que caso o candidato Bush tivesse recebido 152.846 votos em um determinado condado, a votação do canditato Buchanan seria de aproximandamente 791 votos.

Figura 3: Relação entre votos recebidos por Bush e por Bachanan nos condados da Fló



É possível elaborar um código de programação em linguagem R que realize esta mesma previsão utilizando-se do método de Estimativas de Mínimos Quadrados, para se obter $\hat{\beta}_0$ e $\hat{\beta}_1$ e assim, prever um valor de votação do canditato Buchanan sob as mesmas condições, como se segue:

```
x = dados$BUSH
y = dados$BUCHANAN
n = length(x)
beta1 = (sum(x*y) - as.numeric(sum(x)) * as.numeric(sum(y))/n) /
  (sum(x^2) - ((sum (x))^2)/n)
beta0 = mean(y) - beta1 * mean(x)
votos = round(beta0 + 152846 * beta1, 0)

cat("Caso o candidato Bush tivesse recebido 152.846 votos em um
    determinado condado, a votação do canditato Buchanan seria
    de aproximandamente", votos," votos.")
```

Executando o código acima, obtem-se a estimação por mínimos quadrados de 797 votos. Observa-se que por regressão linear a resposta foi de 791 votos, desta forma, percebe-s que ambos os métodos foram similares na predição da votação do candidato Buchanan.

Conclusão

Parte 1.

A atividade permite avaliar um modelos de predição que tem boa aderência aos dados disponibilizados, ao mesmo tempo que mostra a fragilidade da técnica quando assume que a correlação entre as votações dos candidatos pode ser explicada apenas pelos votos observados.

Apêndice

Script em linguagem R

```
#####################################################################
#                                                                   #
#                          SEGUNDA PARTE                            #
#                                                                   #
#####################################################################


# DADOS ----
dados <- read.csv2("florida.csv")
m.bh = max(dados$BUSH)
m.bn = max(dados$BUCHANAN)
aux = dados %>% dplyr::filter(dados$BUSH < m.bh)
dados.sem = aux %>% dplyr::filter(aux$BUCHANAN < m.bn)

# GRÁFICOS DE DISPERSÃO BUBH X BUCHANAN ----

d1 <- dados |>
  ggplot(aes(
    x = BUSH,
    y = BUCHANAN, color = BUCHANAN)) +
  geom_point()+
  labs(
    title = '',
    x = 'Votos em Bush',
    y = 'Votos em Buchanan'
  )+
  scale_y_continuous(
    labels = scales::number_format(
      big.mark = ".",
      decimal.mark = ","
    ))

d1.1 <- dados.sem |>
  ggplot(aes(
    x = BUSH,
    y = BUCHANAN, color = BUCHANAN)) +
  geom_point()+
  labs(
    title = '',
    x = 'Votos em Bush',
```

```
    y = 'Votos em Buchanan'
  )+
  scale_y_continuous(
    labels = scales::number_format(
      big.mark = ".",
      decimal.mark = ","
    ))
d1+d1.1+plot_annotation(
  title = "Figura 1: Relação entre votos recebidos por Bush e
  por Bachanan nos condados da Flórida. (a) Dados completos,
  (b) Sem dados discrepantes",
   caption = "Fonte: 2000 OFFICIAL PRESIDENTIAL GENERAL ELECTION
  RESULTS - USA") &
  theme_bw(base_size = 8) &
  theme(
    legend.position = "none",
    plot.tag.position = c(0, 1),
    plot.tag = element_text(size = 8, hjust = 0, vjust = 0)
  )


# GRÁFICOS DE DISPERSÃO GORE X BUCHANAN ----

d1 <- dados |>
  ggplot(aes(
    x = GORE,
    y = BUCHANAN, color = BUCHANAN)) +
  geom_point()+
  labs(
    title = '',
    x = 'Votos em Gore',
    y = 'Votos em Buchanan'
  )+
  scale_y_continuous(
    labels = scales::number_format(
      big.mark = ".",
      decimal.mark = ","
    ))

d1.1 <- dados.sem |>
  ggplot(aes(
    x = GORE,
    y = BUCHANAN, color = BUCHANAN)) +
  geom_point()+
```

```
  labs(
    title = '',
    x = 'Votos em Gore',
    y = 'Votos em Buchanan'
  )+
  scale_y_continuous(
    labels = scales::number_format(
      big.mark = ".",
      decimal.mark = ","
    ))

d2 <- dados |>
  ggplot(aes(
    x = TOTAL,
    y = BUCHANAN, color = BUCHANAN)) +
  geom_point()+
  labs(
    title = '',
    x = 'Total de votos do condado',
    y = 'Votos em Buchanan'
  )+
  scale_y_continuous(
    labels = scales::number_format(
      big.mark = ".",
      decimal.mark = ","
    ))

d2.2 <- dados.sem |>
  ggplot(aes(
    x = TOTAL,
    y = BUCHANAN, color = BUCHANAN)) +
  geom_point()+
  labs(
    title = '',
    x = 'Total de votos do condado',
    y = 'Votos em Buchanan'
  )+
  scale_y_continuous(
    labels = scales::number_format(
      big.mark = ".",
      decimal.mark = ","
    ))
(d1+d1.1)/(d2+d2.2) + plot_annotation(
  title = "Figura 2: Relação entre votos recebidos por Gore e por
```

```
  Bachanan e com o total de votos nos condados da Flórida.
  (a) Dados completos, (b) Sem dados discrepantes",
  caption = "Fonte: 2000 OFFICIAL PRESIDENTIAL GENERAL ELECTION
  RESULTS - USA") &
  theme_bw(base_size = 8) &
  theme(
    legend.position = "none",
    plot.tag.position = c(0, 1),
    plot.tag = element_text(size = 8, hjust = 0, vjust = 0)
  )

# REGRESSÃO LINEAR ----

reg = lm(dados.sem$BUSH ~ dados.sem$BUCHANAN)
intercepto = reg$coefficients[1][[1]]
c.angular = reg$coefficients[2][[1]]
resposta = round((152846 - intercepto)/c.angular, 0)


d1<- dados.sem |>
  ggplot(aes(
    x = BUSH,
    y = BUCHANAN, color = BUCHANAN)) +
  geom_point()+
  labs(
    title = '',
    x = 'Votos em Bush',
    y = 'Votos em Buchanan'
  )+
  scale_y_continuous(
    labels = scales::number_format(
      big.mark = ".",
      decimal.mark = ","
    ))+
  geom_smooth(method=lm, se=FALSE)

d1+plot_annotation(
  title = "Figura 3: Relação entre votos recebidos por Bush e por
  Bachanan nos condados da Flórida e sua regressão linear.") &
  theme_bw(base_size = 8) &
  theme(
    legend.position = "none",
    plot.tag.position = c(0, 1),
    plot.tag = element_text(size = 8, hjust = 0, vjust = 0)
```

```
  )

# MÍNIMOS QUADRADOS ----

x = dados$BUSH
y = dados$BUCHANAN
n = length(x)
beta1 = (sum(x*y) - as.numeric(sum(x)) * as.numeric(sum(y))/n) /
  (sum(x^2) - ((sum (x))^2)/n)
beta0 = mean(y) - beta1 * mean(x)
votos = round(beta0 + 152846 * beta1, 0)

cat("Caso o candidato Bush tivesse recebido 152.846 votos em um
    determinado condado, a votação do canditato Buchanan seria
    de aproximandamente", votos," votos.")
```