Lição 12: Regressão Linear Parte 1

N.B.: Rode as linhas de comando a seguir antes de iniciar esta lição. Defina como diretório de trabalho aquele que contém o arquivo Pretonicas.csv.

```
# Definir diretório de trabalho

#setwd()

# Importar planilha de dados

pretonicas <- read_csv("Pretonicas.csv",
                       col_types = cols(.default = col_factor(),
                                        VOGAL = col_factor(levels = c(
"i", "e", "a", "o", "u")),
                                        F1 = col_double(),
                                        F2 = col_double(),
                                        F1.NORM = col_double(),
                                        F2.NORM = col_double(),
                                        F1.SIL.SEG = col_double(),
                                        F2.SIL.SEG = col_double(),
                                        F1.SEG.NORM = col_double(),
                                        F2.SEG.NORM = col_double(),
                                        DIST.TONICA = col_double(),
                                        Begin.Time.s = col_double(),
                                        End.Time.s = col_double(),
                                        Duration.ms = col_double(),
                                        IDADE = col_integer(),
                                        IDADE.CHEGADA = col_integer(),
                                        ANOS.SP = col_integer()
                                        )
                       )

pretonicas$CONT.PREC <-  fct_collapse(pretonicas$CONT.PREC,
                          dental.alveolar = c("t", "d", "n", "l"),
                          labial = c("p", "b", "m", "f", "v"),
                          palatal.sibilante = c("S", "Z", "L", "s", "z
"),
                          velar = c("k", "g"),
                          vibrante = c("h", "R")
                                    )

pretonicas$CONT.PREC <- fct_relevel(pretonicas$CONT.PREC, "dental.alve
olar", "labial", "palatal.sibilante", "velar", "vibrante")

pretonicas$CONT.SEG <- fct_collapse(pretonicas$CONT.SEG,
                     dental.alveolar = c("t", "d", "n", "l"),
                     labial = c("p", "b", "m", "f", "v"),
                     palatal.sibilante = c("S", "Z", "L", "N", "s", "z
```

```
"),
                      velar = c("k", "g"),
                      vibrante = c("r", "h", "R")
                                  )

pretonicas$CONT.SEG <- fct_relevel(pretonicas$CONT.SEG, "dental.alveol
ar", "labial", "palatal.sibilante", "velar", "vibrante")

### Criar subconjunto de dados da vogal /e/ pretonica

VOGAL_e <- filter(pretonicas, VOGAL == "e") %>%
  droplevels()
```

Nas Lições 10 e 11, vimos testes estatísticos que se aplicam a variáveis dependentes numéricas: o teste-t e sua versão não paramétrica podem ser usados quando se quer comparar as distribuições de dois grupos – ou seja, aplicam-se quando se tem uma VD numérica e uma VI nominal binária (ou uma distribuição conhecida); o teste de correlação, por sua vez, aplica-se quando se tem uma VD numérica e uma VI também numérica. Esses testes só permitem testar correlações entre uma VD e uma única VI – são análises *univariadas*. Nesta e nas próximas aulas veremos quais análises podem ser feitas quando se trabalha com mais de uma variável independente – *modelos multivariados*.

Na última lição, também vimos que uma função linear pode ser denotada pela expressão y = a + bx, em que "a" é o coeficiente linear e "b" é o coeficiente angular. Análises multivariadas seguem uma estrutura semelhante, mas que inclui novos coeficientes angulares, um para cada variável independente no modelo: y = a + bx1 + cx2 + dx3... b, c e d, nesse exemplo, são coeficientes angulares das variáveis x1, x2 e x3. Em análises multivariadas, a variável dependente y é chamada de *variável resposta*, e as variáveis independentes x1, x2 etc. são chamadas de *variáveis previsoras*. Mais adiante, veremos por que os termos "variável dependente" e "variável independente" não são adequados para modelos de regressão.

Para começar, carregue o pacote tidyverse.

```
library(tidyverse)
```

Carregue também o pacote effects, que usaremos para visualizar resultados de modelos de regressão.

```
library(effects)
```

Vamos trabalhar aqui com os dados da vogal /e/ pretônica, na fala de migrantes paraibanos residentes em São Paulo e na fala de paulistanos nativos. No *script*, deixei essas linhas de comando, mas você não precisará rodá-las (ficam apenas para você saber como o dataframe foi criado). Esses dados foram guardados no df `VOGAL_e`. Aplique a função `str()` para inspecioná-lo. Em especial, veja as variáveis AMOSTRA, SEXO, F1.SEG.NORM, CONT.PREC e CONT.SEG, com as quais trabalharemos nesta e na próxima lição.

```
str(VOGAL_e)
```

```
## spec_tbl_df [686 × 27] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ PALAVRA      : Factor w/ 259 levels “diferente”,”melhor”,..: 1 1
2 3 1 4 5 6 7 8 ...
##  $ Transc.Fon   : Factor w/ 259 levels “d<i>-f<e>-'RE-te”,..: 1 1 2
3 1 4 5 6 7 8 ...
##  $ VOGAL        : Factor w/ 1 level “e”: 1 1 1 1 1 1 1 1 1 1 ...
##  $ F1           : num [1:686] 613 656 573 735 567 ...
##  $ F2           : num [1:686] 2014 1848 2413 1656 1375 ...
##  $ F1.NORM      : num [1:686] 447 464 431 496 429 ...
##  $ F2.NORM      : num [1:686] 1698 1611 1905 1512 1366 ...
##  $ CONT.PREC    : Factor w/ 5 levels “dental.alveolar”,..: 2 2 2 5
2 4 2 3 2 1 ...
##  $ CONT.SEG     : Factor w/ 5 levels “dental.alveolar”,..: 5 5 3 2
5 5 1 5 3 2 ...
##  $ VOGAL.SIL.SEG: Factor w/ 11 levels “a”,”aw”,”A”,”\u0097”,..: 5 5
4 7 5 5 1 5 1 5 ...
##  $ F1.SIL.SEG   : num [1:686] 569 524 686 652 661 ...
##  $ F2.SIL.SEG   : num [1:686] 1674 2428 1497 2159 1865 ...
##  $ F1.SEG.NORM  : num [1:686] 350 336 385 375 378 ...
##  $ F2.SEG.NORM  : num [1:686] 1360 1724 1274 1594 1452 ...
##  $ VOGAL.TONICA : Factor w/ 14 levels “e”,”o”,”ow”,”a”,..: 9 9 2 1
9 9 4 9 4 9 ...
##  $ DIST.TONICA  : num [1:686] 1 1 1 1 1 1 1 1 1 2 ...
##  $ ESTR.SIL.PRET: Factor w/ 5 levels “CV”,”CVs”,”CCV”,..: 1 1 1 1 1
1 1 1 1 1 ...
##  $ Begin.Time.s : num [1:686] 219 226 576 584 614 ...
##  $ End.Time.s   : num [1:686] 219 226 576 584 614 ...
##  $ Duration.ms  : num [1:686] 10.4 11.6 30.3 17.5 17.1 ...
##  $ AMOSTRA      : Factor w/ 2 levels “PBSP”,”SP2010”: 1 1 1 1 1 1 1
1 1 1 ...
##  $ PARTICIPANTE : Factor w/ 14 levels “MartaS”,”JosaneV”,..: 1 1 1
1 1 1 1 1 1 1 ...
##  $ SEXO         : Factor w/ 2 levels “feminino”,”masculino”: 1 1 1
1 1 1 1 1 1 1 ...
##  $ IDADE        : int [1:686] 32 32 32 32 32 32 32 32 32 32 ...
##  $ IDADE.CHEGADA: int [1:686] 18 18 18 18 18 18 18 18 18 18 ...
##  $ ANOS.SP      : int [1:686] 14 14 14 14 14 14 14 14 14 14 ...
```

```
##  $ CONTEXTO     : Factor w/ 632 levels "diferente o clima de eu com
ele",..: 1 2 3 4 5 6 7 8 9 10 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   .default = col_factor(),
##   ..   PALAVRA = col_factor(levels = NULL, ordered = FALSE, include
_na = FALSE),
##   ..   Transc.Fon = col_factor(levels = NULL, ordered = FALSE, incl
ude_na = FALSE),
##   ..   VOGAL = col_factor(levels = c("i", "e", "a", "o", "u"), orde
red = FALSE, include_na = FALSE),
##   ..   F1 = col_double(),
##   ..   F2 = col_double(),
##   ..   F1.NORM = col_double(),
##   ..   F2.NORM = col_double(),
##   ..   CONT.PREC = col_factor(levels = NULL, ordered = FALSE, inclu
de_na = FALSE),
##   ..   CONT.SEG = col_factor(levels = NULL, ordered = FALSE, includ
e_na = FALSE),
##   ..   VOGAL.SIL.SEG = col_factor(levels = NULL, ordered = FALSE, i
nclude_na = FALSE),
##   ..   F1.SIL.SEG = col_double(),
##   ..   F2.SIL.SEG = col_double(),
##   ..   F1.SEG.NORM = col_double(),
##   ..   F2.SEG.NORM = col_double(),
##   ..   VOGAL.TONICA = col_factor(levels = NULL, ordered = FALSE, in
clude_na = FALSE),
##   ..   DIST.TONICA = col_double(),
##   ..   ESTR.SIL.PRET = col_factor(levels = NULL, ordered = FALSE, i
nclude_na = FALSE),
##   ..   Begin.Time.s = col_double(),
##   ..   End.Time.s = col_double(),
##   ..   Duration.ms = col_double(),
##   ..   AMOSTRA = col_factor(levels = NULL, ordered = FALSE, include
_na = FALSE),
##   ..   PARTICIPANTE = col_factor(levels = NULL, ordered = FALSE, in
clude_na = FALSE),
##   ..   SEXO = col_factor(levels = NULL, ordered = FALSE, include_na
= FALSE),
##   ..   IDADE = col_integer(),
##   ..   IDADE.CHEGADA = col_integer(),
##   ..   ANOS.SP = col_integer(),
##   ..   CONTEXTO = col_factor(levels = NULL, ordered = FALSE, includ
e_na = FALSE)
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Aplique também a função `View()` para se (re)familiarizar com a planilha.

```
View(VOGAL_e)
```

N.B.: Resultado aqui omitido.

A variável AMOSTRA é codificada de acordo com a amostra da qual vieram os dados: PBSP (migrantes paraibanos) ou SP2010 (paulistanos nativos). SEXO indica se o falante é do sexo feminino ou masculino. A variável F1.SEG.NORM contém as medições normalizadas da altura (F1) da vogal da sílaba seguinte – por exemplo, em "relógio", refere-se à medida de F1 da vogal /ɔ/. CONT.PREC e CONT.SEG codificam o segmento fonológico que precede ou antecede a vogal pretônica.

É importante reforçar que, antes de chegar ao ponto de análises multivariadas, idealmente o pesquisador já terá feito análises preliminares, por meio de tabelas e gráficos, e já terá uma boa ideia de como se dá a distribuição de seus dados: se há poucas ocorrências em certos fatores, se parece haver diferenças entre fatores de uma variável, se os testes univariados apontam para diferenças significativas ou não. Por outro lado, se o pesquisador está trabalhando com mais de uma VI/variável previsora, é *imprescindível* realizar análises multivariadas, pois o comportamento de certas variáveis pode não ser tão preponderante ou pode mudar em face de outras.

A função empregada para criar modelos multivariados para uma variável dependente numérica é `lm()`, que já vimos na Lição 11. Aqui vamos aplicá-la à variável resposta numérica altura da vogal /e/, em Hz: F1.NORM. Nesta lição, vamos criar modelos com apenas uma variável previsora a fim de treinar a leitura dos resultados. Contudo, o interesse principal nessa função é a modelagem multivariada – que faremos na próxima.

Em outras palavras, as tarefas desta lição têm objetivo puramente didático; as análises que você fará efetivamente com seus dados e que acabará reportando serão mais parecidas com o que faremos na Lição 13.

Crie um primeiro modelo para testar se há correlação entre a altura da vogal e a AMOSTRA, ou seja, para testar se a altura da vogal /e/ na fala de paraibanos e de paulistanos difere. O primeiro argumento de `lm()` é uma fórmula y ~ x, e o segundo argumento é o conjunto de dados. Digite então a linha de comando a seguir para criar o modelo `mod1`.

```
mod1 <- lm(F1.NORM ~ AMOSTRA, data = VOGAL_e)
```

Visualize agora o resumo do resultado de `mod1` com `summary()`.

```
summary(mod1)

##
## Call:
## lm(formula = F1.NORM ~ AMOSTRA, data = VOGAL_e)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -75.71 -18.31  -2.20  16.12 145.56
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    431.756      1.480 291.767  < 2e-16 ***
## AMOSTRASP2010   -8.790      2.084  -4.219 2.79e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.29 on 684 degrees of freedom
## Multiple R-squared:  0.02536,    Adjusted R-squared:  0.02394
## F-statistic:  17.8 on 1 and 684 DF,  p-value: 2.788e-05
```

Vamos examinar o resultado do modelo linear. O primeiro ponto a se checar é se os resíduos têm distribuição normal: valor de mediana próximo a zero e valores de mínimo-máximo e 1Q-3Q razoavelmente simétricos. Olhando os resíduos de `mod1`, qual é a sua avaliação?

- a distribuição dos resíduos segue a distribuição normal
- a distribuição dos resíduos não segue a distribuição normal

O valor máximo de resíduo, 145,56, difere muito do valor mínimo, -75,71 em valores absolutos. Isso é indicativo de que há valores atípicos na distribuição. Lembre-se que os resíduos são a diferença entre os valores observados e os valores previstos pelo modelo; um resíduo tão grande, de 145 Hz, refere-se a alguma observação muito acima do esperado.

Visualize a distribuição das medidas de F1.NORM por AMOSTRA nos dados `VOGAL_e`, por meio de um boxplot (Figura 12.1).

```
ggplot(VOGAL_e, aes(x = AMOSTRA, y = F1.NORM)) +
  geom_boxplot(notch = TRUE) +
  scale_y_reverse()
```

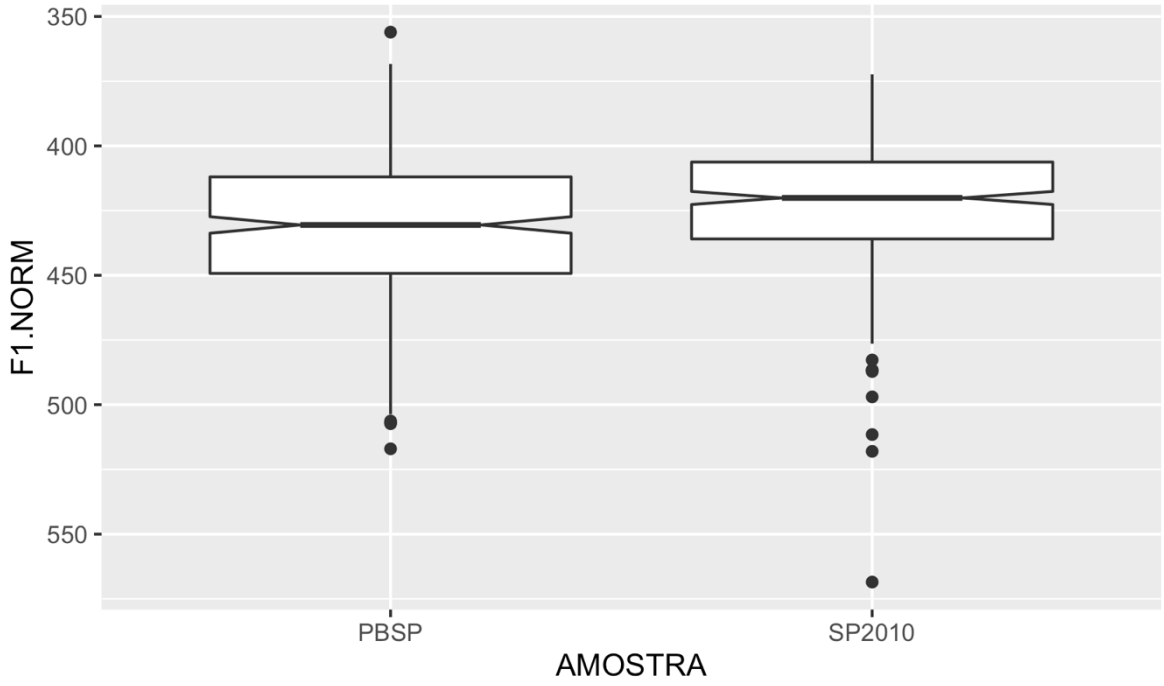

Figura 12.1: Boxplots das medidas de F1 normalizado da vogal /e/ nas amostras PBSP e SP2010. Fonte: própria.

Vemos que as medidas de F1.NORM se concentram entre 400 e 450 Hz, e que há alguns poucos valores atípicos acima de 500 Hz. Crie então um novo subconjunto de dados, chamado `VOGAL_e2`, que inclui os dados de `VOGAL_e` cuja medida de F1.NORM é abaixo de 500 Hz.

```
VOGAL_e2 <- filter(VOGAL_e, F1.NORM < 500)
```

Faça um novo modelo linear, chamado `mod2`, com a mesma fórmula de `mod1` e os dados de `VOGAL_e2`.

```
mod2 <- lm(F1.NORM ~ AMOSTRA, data = VOGAL_e2)
```

Visualize o resumo do resultado de `mod2`.

```
summary(mod2)

##
## Call:
## lm(formula = F1.NORM ~ AMOSTRA, data = VOGAL_e2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -74.357 -17.225  -1.261  16.848  74.960
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     430.406      1.391 309.423  < 2e-16 ***
## AMOSTRASP2010   -8.400      1.954  -4.298 1.97e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.42 on 675 degrees of freedom
## Multiple R-squared:  0.02664,    Adjusted R-squared:  0.0252
## F-statistic: 18.48 on 1 and 675 DF,  p-value: 1.975e-05
```

Comparativamente a `mod1`, os resíduos de `mod2` são mais simétricos: os valores absolutos de mínimo e máximo são em torno de 74; os valores absolutos de 1Q-3Q são próximos de 17, e a mediana está razoavelmente próxima de zero.

Na sequência do resultado, temos a tabela de coeficientes, com as estimativas, erro padrão, valor-*t* e significância. Assim como no modelo linear que vimos na Lição 11, a significância aqui testa a hipótese nula de que a estimativa é igual a zero. Pela tabela, ambas as estimativas diferem significativamente de zero. Como interpretar esses números?

Quando testamos a correlação entre duas variáveis numéricas (Lição 11), o coeficiente linear mediu a estimativa de y (a VD/variável resposta) quando x (a VI/variável previsora) = 0. Aqui, x se refere à variável AMOSTRA. O que significa AMOSTRA = 0? Trata-se do valor de referência da variável, que é seu primeiro nível (na ordem da tabela ou modificado pelo usuário). Nesse df, PBSP é o primeiro nível e, portanto, a estimativa de 430,4 Hz representa o valor médio esperado de F1.NORM para a vogal /e/ dos paraibanos em São Paulo. O valor-*p* abaixo de 0,05, para nós, nada significa aqui; a medida apenas indica que é baixa a probabilidade de se ter observado 430 caso o verdadeiro parâmetro seja zero, mas nunca esperaríamos que uma vogal /e/ tivesse 0 Hz de F1!

A estimativa para SP2010, -8,4 Hz, deve ser lida em relação ao coeficiente linear. O modelo estima que o valor de F1.NORM para paulistanos é 430,4 - 8,4 = 422 Hz – ou seja, em média, os valores de F1.NORM para paulistanos são menores, o que indica que a vogal pretônica /e/ de paulistanos tende a ser significativamente mais alta. Os resultados

de `mod2` podem ser colocados na forma da função F1.NORM = 430,4 - 8,4 * AMOSTRASP2010.

Se você se lembra do teste-t da Lição 10, vimos valores semelhantes para as médias de F1.NORM para cada amostra (430 Hz para PBSP e 422 Hz para SP2010). Há uma pequena diferença, que se deve ao fato de que aqui excluímos alguns valores atípicos para `mod2`. Faça esse teste: aplique um teste-t para testar a hipótese nula de que a diferença entre as médias de F1.NORM entre paraibanos e paulistanos é zero, no subconjunto de dados `VOGAL_e2`.

```
t.test(F1.NORM ~ AMOSTRA, data = VOGAL_e2)
```

```
##
##  Welch Two Sample t-test
##
## data:  F1.NORM by AMOSTRA
## t = 4.2863, df = 639.61, p-value = 2.097e-05
## alternative hypothesis: true difference in means between group PBSP
and group SP2010 is not equal to 0
## 95 percent confidence interval:
##   4.551609 12.247925
## sample estimates:
##   mean in group PBSP mean in group SP2010
##             430.4061             422.0064
```

As médias calculadas no teste-t correspondem exatamente às estimativas do modelo linear. Na prática, ao incluir apenas uma VI na função `lm()`, estamos efetivamente executando uma análise univariada como o teste-t. A diferença do modelo linear, nesse caso, é visualizar o resultado da estimativa em termos da *diferença* entre os níveis. Não subestime esse fato, pois isso é uma *grande* vantagem: isso torna mais fácil verificar se essa diferença é ou não zero.

As demais informações são interpretadas como já comentado na lição anterior: o erro padrão dos resíduos indica o quanto da variação o modelo não é capaz de explicar e R^2/R^2-ajustado indicam o quanto da variação nos dados é explicada pelas variáveis incluídas no modelo. A estatística-F permite avaliar a significância do modelo como um todo e comparar diferentes modelos.

O pacote effects permite plotar gráficos de efeitos a partir de modelos lineares e logísticos criados no R. Aplique a função `plot()` com os seguintes argumentos: (i) para

os dados a plotar, digite `effect("AMOSTRA", mod2)`; (ii) para os limites do eixo y, coloque ylim = c(450, 400) – para que os valores do eixo fiquem invertidos; e (iii) inclua o argumento grid = T (Figura 12.2).

```
plot(effect("AMOSTRA", mod2), ylim = c(450, 400), grid = T)
```

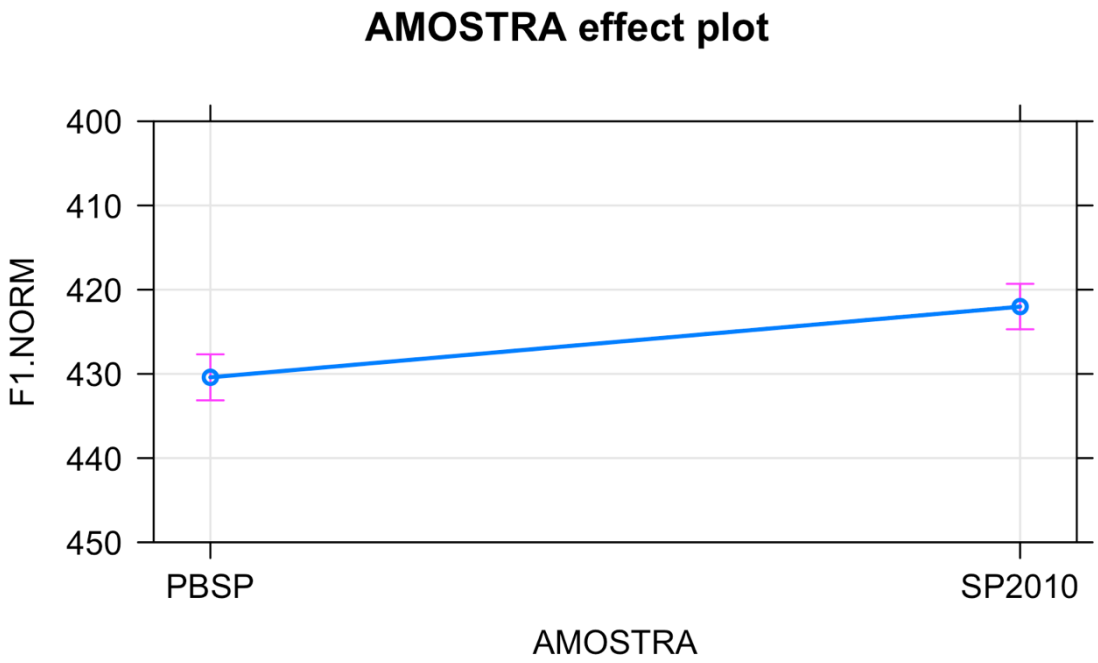

Figura 12.2: Gráfico de efeitos da variável Amostra para a altura da vogal /e/ pretônica. Fonte: própria.

O gráfico de efeitos plota as estimativas de cada um dos níveis da variável indicada dentro da função `effect()`, extraídas do modelo, junto com o intervalo de confiança indicado pelas barras. Vê-se que a diferença entre os níveis da variável AMOSTRA é significativa porque os intervalos não se sobrepõem: mesmo que a medida de F1.NORM para PBSP fosse mais alta, e a medida de F1.NORM para SP2010 fosse mais baixa, elas ainda assim não seriam iguais dentro do intervalo de confiança de 95% – o que significa que é menor do que 5% a probabilidade de que as médias sejam iguais, que não há diferença entre as amostras.

Aqui usamos a função `plot()`, da instalação base do R, para plotar esse gráfico. Também é possível fazer isso com o ggplot2, mas aí são necessários outros passos. Para sua curiosidade, deixei as linhas de comando prontas ao final do *script* dessa lição. Rode-

as posteriormente para ver como ficaria um gráfico no ggplot2 (como sempre, mais bonito do que com a função da instalação base!).

Vejamos agora um modelo que inclui uma variável previsora com mais de dois níveis. Faça um modelo linear, chamado `mod3`, com a variável resposta F1.NORM e a variável previsora CONT.SEG, nos dados `VOGAL_e2`.

```
mod3 <- lm(F1.NORM ~ CONT.SEG, data = VOGAL_e2)
```

Veja o resultado de `mod3` com a função `summary()`.

```
summary(mod3)

##
## Call:
## lm(formula = F1.NORM ~ CONT.SEG, data = VOGAL_e2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -72.167 -17.788  -1.421  15.539  73.615
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 433.429      3.007 144.162  < 2e-16 ***
## CONT.SEGlabial              -11.106      3.943  -2.817 0.004997 **
## CONT.SEGpalatal.sibilante   -14.299      4.014  -3.562 0.000394 ***
## CONT.SEGvelar                -7.108      3.546  -2.004 0.045425 *
## CONT.SEGvibrante             -5.213      3.447  -1.512 0.130906
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.51 on 672 degrees of freedom
## Multiple R-squared:  0.02409,    Adjusted R-squared:  0.01828
## F-statistic: 4.147 on 4 and 672 DF,  p-value: 0.002515
```

Vamos direto aos coeficientes desta vez. Para bem entender o que significam as estimativas, é necessário saber quais são os níveis da variável CONT.SEG. Digite `levels(VOGAL_e2$CONT.SEG)`.

```
levels(VOGAL_e2$CONT.SEG)

## [1] “dental.alveolar”   “labial”           “palatal.sibilante”
## [4] “velar”             “vibrante”
```

A partir dos níveis de CONT.SEG, a que se refere a estimativa do Intercept?

- dental.alveolar
- labial
- vibrante

- velar
- palatal.sibilante

O valor de 433,4 Hz, portanto, é a estimativa da medida de F1.NORM quando a consoante seguinte é uma dental-alveolar (p.ex. pe*d*agogia). Os demais valores devem ser lidos em relação a esse nível, somando-se o valor das respectivas estimativas ao coeficiente linear (Intercept). Assim, a estimativa da medida de F1.NORM para quando a consoante seguinte é labial (p.ex., de*m*ora) é 433,4 - 11,1 = 422,3 Hz; para as consoantes palatais ou sibilantes (p.ex. pe*s*ado) é 433,4 - 14,3 = 419,1 Hz etc. (lembre-se que a soma de um valor negativo equivale à subtração!). Em outras palavras, a estimativa de cada termo previsor é computada pela função F1.NORM = 433,429 + (-11,106 * CONT.SEGlabial) + (-14,299 * CONT.SEGpalatal.sibilante) + (-7,108 * CONT.SEGvelar) + (-5,213 * CONT.SEGvibrante).

Em relação a consoantes dental-alveolares, quais níveis têm medidas de F1.NORM significativamente diferentes?

- segmentos palatais-sibilantes
- segmentos palatais e vibrantes
- segmentos labiais, palatais-sibilantes e velares

Em relação a consoantes dental-alveolares, quais níveis não têm medidas de F1.NORM significativamente diferentes?

- segmentos vibrantes
- segmentos palatais-sibilantes
- segmentos labiais, palatais-sibilantes e velares

Façamos agora um gráfico de efeitos para visualizar as diferenças entre as consoantes. A partir da linha de comando com `plot()` e `effect()` digitada acima, substitua o nome da variável para CONT.SEG e do modelo para `mod3` (Figura 12.3).

```
plot(effect("CONT.SEG", mod3), ylim = c(450, 400), grid = T)
```

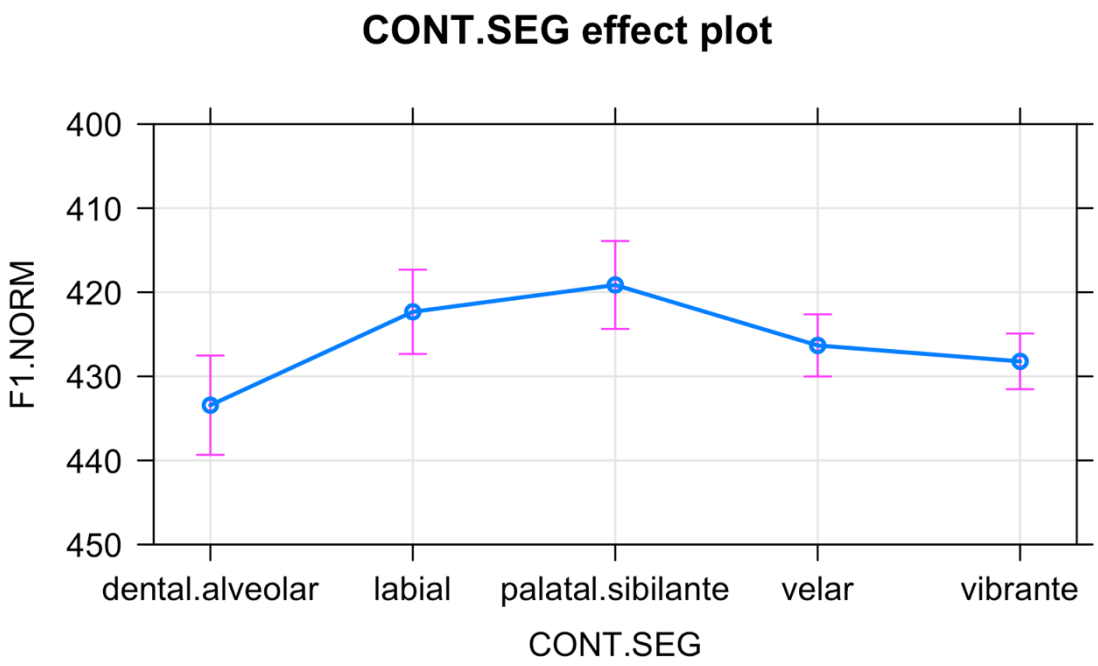

Figura 12.3: Gráfico de efeitos da variável Contexto Fônico Seguinte para a altura da vogal /e/ pretônica. Fonte: própria.

No gráfico de efeitos, vemos que as labiais, palatais-sibilantes e velares se distanciam mais da média de F1.NORM em relação às dental-alveolares. Como estas são o valor de referência de CONT.SEG, as estimativas de todos os níveis são calculados a partir desse nível. Mas e se quiséssemos saber se há diferenças significativas na medida de F1.NORM entre, por exemplo, vibrantes e palatais-sibilantes? ou entre labiais e velares? Seria necessário mudar o nível de referência a cada novo teste?

Para múltiplas comparações, pode-se usar o método de Tukey, por meio da função `TukeyHSD()`. Digite `TukeyHSD(aov(mod3))` para verificar o teste de significância de todos os pares possíveis dos níveis de CONT.SEG.

```
TukeyHSD(aov(mod3))

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = mod3)
##
## $CONT.SEG
##                                          diff         lwr        upr
## labial-dental.alveolar            -11.105787 -21.8908147 -0.3207598
## palatal.sibilante-dental.alveolar -14.299162 -25.2787251 -3.3195998
```

```
## velar-dental.alveolar                -7.108380 -16.8082987  2.5915390
## vibrante-dental.alveolar             -5.213098 -14.6411626  4.2149660
## palatal.sibilante-labial             -3.193375 -13.2738266  6.8870762
## velar-labial                          3.997407  -4.6716831 12.6664979
## vibrante-labial                       5.892689  -2.4711082 14.2564860
## velar-palatal.sibilante               7.190783  -1.7191618 16.1007270
## vibrante-palatal.sibilante            9.086064   0.4728719 17.6992564
## vibrante-velar                        1.895282  -5.0130258  8.8035889
##                                         p adj
## labial-dental.alveolar              0.0399172
## palatal.sibilante-dental.alveolar 0.0036031
## velar-dental.alveolar               0.2649058
## vibrante-dental.alveolar            0.5547385
## palatal.sibilante-labial            0.9090578
## velar-labial                        0.7151076
## vibrante-labial                     0.3038915
## velar-palatal.sibilante             0.1781207
## vibrante-palatal.sibilante          0.0327983
## vibrante-velar                      0.9443750
```

O resultado do teste de Tukey é uma tabela com as estimativas de diferença entre cada par, o intervalo de confiança (os limites lwr e upr) e o valor-*p* ajustado (já que são múltiplas comparações). Veja que só são significativas as diferenças cujo intervalo de confiança não inclui zero. A partir dessas comparações, o pesquisador pode decidir juntar novos níveis em um mesmo fator, contanto que também haja justificativa teórica para tal.

Vamos agora fazer um último modelo univariado com uma variável previsora numérica. Crie um modelo chamado `mod4` que testa se há correlação entre F1.NORM e F1.SEG.NORM, a altura da vogal da sílaba seguinte, nos dados de `VOGAL_e2`.

```
mod4 <- lm(F1.NORM ~ F1.SEG.NORM, data = VOGAL_e2)
```

Veja o resultado de `mod4` com `summary()`.

```
summary(mod4)
```

```
##
## Call:
## lm(formula = F1.NORM ~ F1.SEG.NORM, data = VOGAL_e2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.456 -18.508  -0.944  15.235  74.681
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 379.50710    9.52279  39.852  < 2e-16 ***
## F1.SEG.NORM   0.12278    0.02494   4.924 1.07e-06 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.32 on 675 degrees of freedom
## Multiple R-squared:  0.03467,    Adjusted R-squared:  0.03324
## F-statistic: 24.24 on 1 and 675 DF,  p-value: 1.068e-06
```

Vamos novamente direto aos coeficientes, para treinar sua leitura. O coeficiente linear deste modelo é 379,5 Hz e o coeficiente angular é 0,12 Hz, e ambos diferem significativamente de zero. Novamente, o fato de o primeiro coeficiente diferir de zero não nos diz nada, pois não esperaríamos que a vogal /e/ tivesse 0 Hz, mas o segundo coeficiente nos traz uma informação relevante: existe uma correlação entre a altura da vogal seguinte e a altura da vogal /e/ pretônica. O coeficiente angular é positivo, o que indica que quanto maior o valor de F1.SEG.NORM, maior o valor de F1.NORM; em termos numéricos, a leitura desse resultado é que a cada unidade de F1.SEG.NORM (400 Hz, 401 Hz, 402 Hz...), estima-se que o valor de F1.NORM aumente em 0,12 Hz. Ou, ainda, a estimativa da variável resposta segue a função F1.NORM = 379,50710 + (0,12278 * F1.SEG.NORM). Digamos que uma vogal da sílaba seguinte à pretônica tenha F1.SEG.NORM de 400 Hz; a estimativa do valor de F1.NORM é 379,50710 + (0,12278 * 400) = 428,6191 Hz.

O gráfico de efeitos permite visualizar essa correlação de forma mais clara. A partir da linha de comando em que usamos a função `plot()` e `effect()`, substitua o nome da variável para F1.SEG.NORM e o modelo para `mod4` (Figura 12.4).

```
plot(effect("F1.SEG.NORM", mod4), ylim = c(450, 400), grid = T)
```

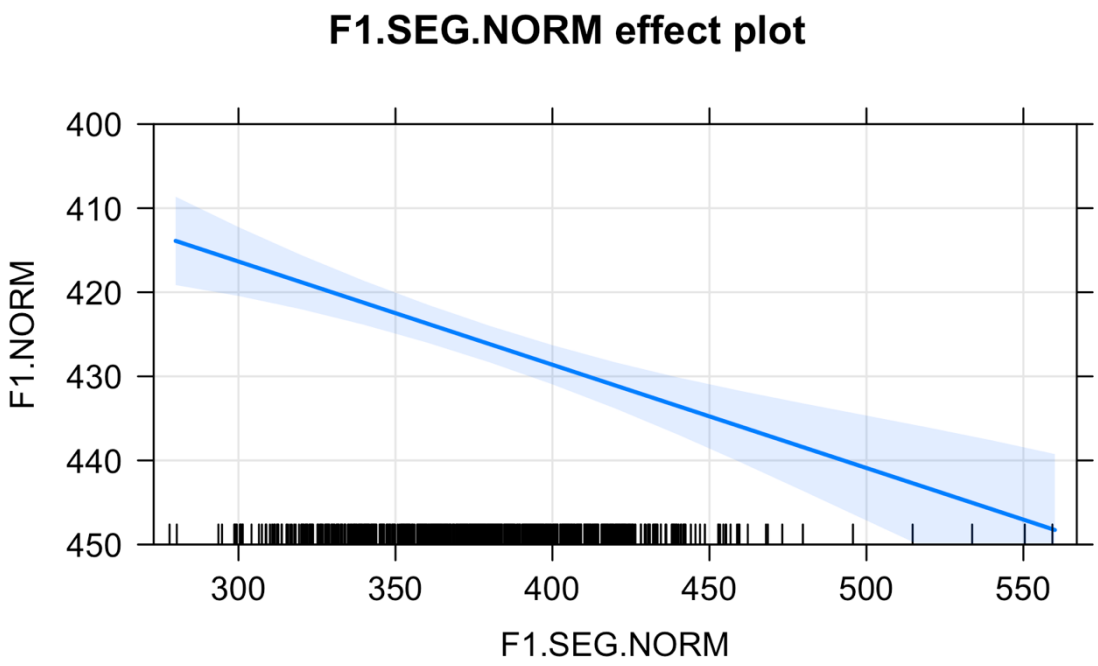

Figura 12.4: Gráfico de efeitos da variável F1 da vogal da sílaba seguinte para a altura da vogal /e/ pretônica. Fonte: própria.

Como F1.SEG.NORM é uma variável numérica, o tipo de gráfico mudou para uma linha de regressão. A figura mostra uma curva descendente porque invertemos o eixo y para adequá-lo à convenção de representar valores mais altos de F1 na parte de baixo. Aí vemos claramente que quanto maior o valor de F1.SEG.NORM, mais baixa tende a ser a vogal /e/ pretônica. (Eis aí o fenômeno de harmonia vocálica!) A mancha em volta da linha de regressão indica o intervalo de confiança das estimativas. Os pequenos traços verticais ao longo do eixo x mostram onde estão e onde se concentram as observações: há muito mais dados de F1.SEG.NORM entre cerca de 330 Hz e 420 Hz. Veja que o intervalo de confiança das estimativas se "alarga" nas partes em que há menor número de observações, justamente porque é mais difícil chegar a estimativas precisas quando não temos muitos dados.

Vimos então como ler os resultados de um modelo linear com uma variável previsora binária (`mod1` e `mod2`), uma variável previsora com mais de 2 fatores (`mod3`) e uma variável previsora numérica (`mod4`). Podemos agora partir para modelos um pouco mais complexos. Façamos então um modelo `mod5` com a inclusão de duas variáveis

previsoras, AMOSTRA + SEXO. Digite `mod5 <- lm(F1.NORM ~ AMOSTRA + SEXO, data = VOGAL_e2)`.

```
mod5 <- lm(F1.NORM ~ AMOSTRA + SEXO, data = VOGAL_e2)
```

E veja o resultado com `summary()`.

```
summary(mod5)
```

```
##
## Call:
## lm(formula = F1.NORM ~ AMOSTRA + SEXO, data = VOGAL_e2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.763 -16.827  -1.507  16.510  74.168
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    431.190      1.789 241.086  < 2e-16 ***
## AMOSTRASP2010   -8.392      1.955  -4.293 2.02e-05 ***
## SEXOmasculino   -1.378      1.975  -0.698    0.486
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.43 on 674 degrees of freedom
## Multiple R-squared:  0.02734,    Adjusted R-squared:  0.02446
## F-statistic: 9.474 on 2 and 674 DF,  p-value: 8.756e-05
```

O modelo agora fornece o valor do coeficiente linear, 431,190 Hz, e dois coeficientes angulares, um para AMOSTRASP2010 e outro para SEXOmasculino. Como o modelo inclui duas variáveis previsoras, o valor do coeficiente linear deve ser interpretado em relação a ambos os níveis de referência: AMOSTRA PBSP e SEXO feminino. O valor de 431,190 Hz, portanto, é a estimativa da altura da vogal /e/ pretônica na fala de *paraibanas*.

Se somarmos a estimativa de AMOSTRASP2010 ao coeficiente linear (431,190 - 8,392 = 422,798 Hz), teremos o valor estimado de F1.NORM para falantes da amostra SP2010 do sexo *feminino* – isso porque, em relação ao nível de referência, alteramos apenas AMOSTRA, não SEXO. Se somarmos a estimativa de SEXOmasculino ao coeficiente linear (431,190 - 1,378 = 429,82 Hz), teremos o valor estimado de F1.NORM para falantes do sexo masculino da AMOSTRA PBSP – isso porque, em relação ao nível de referência, neste caso alteramos apenas SEXO, não AMOSTRA. Para obter a estimativa de F1.NORM para

falantes do sexo masculino da amostra SP2010, é necessário somar ambos os coeficientes angulares ao coeficiente linear: 431,190 - 8,392 - 1,378 = 421,42 Hz (Tabela 12.1).

Tabela 12.1: Cálculo das probabilidades em logodds a partir das estimativas geradas pelo modelo de regressão linear.

	F	M
PBSP	431,190	429,812
SP2010	422,798	421,420

Fonte: própria.

O modelo nos informa que existe uma diferença significativa na altura da vogal /e/ pretônica na fala de paraibanos e paulistanos, mas não há diferença significativa entre os sexos. Isso pode ser mais bem visualizado por meio de um gráfico de efeitos! A partir da última linha de comando em que você empregou a função `plot()`, modifique o primeiro argumento – onde estava effect("F1.SEG.NORM", mod4) – para allEffects(mod5). Em modelos multivariados, em vez de `effect()`, usamos a função `allEffects()` com o modelo como único argumento. Certifique-se de que sua linha de comando é `plot(allEffects(mod5), ylim = c(450, 400), grid = T)` (Figura 12.5).

```
plot(allEffects(mod5), ylim = c(450, 400), grid = T)
```

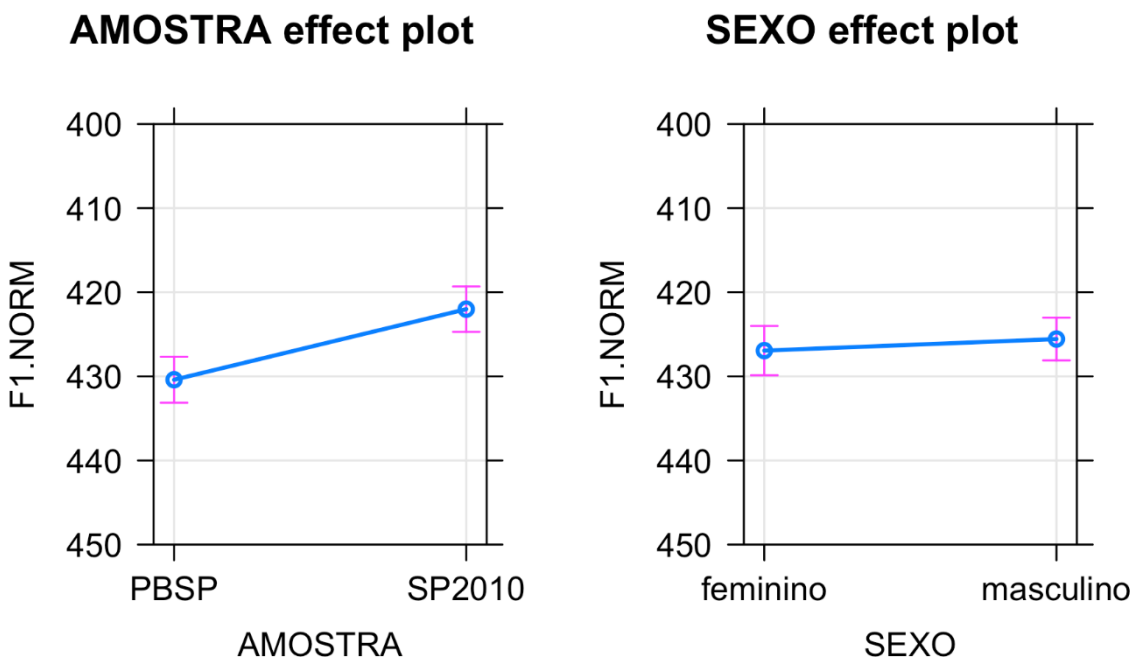

Figura 12.5: Gráfico de efeitos das variáveis Amostra e Sexo para a altura da vogal /e/ pretônica. Fonte: própria.

A Figura 12.5 mostra dois gráficos de efeitos, um para cada variável. Aí vemos que a diferença na altura da vogal /e/ pretônica é significativa para AMOSTRA (pois os intervalos de confiança não se sobrepõem), mas não é para a variável SEXO.

No modelo acima, incluímos duas variáveis previsoras por meio do operador de soma +. Essa notação não é por acaso: existe aí a pressuposição de que o efeito dos previsores é *aditivo*, de modo que, para chegar aos valores estimados de cada uma das combinações possíveis (mulheres paraibanas, homens paraibanos, mulheres paulistanas e homens paulistanos), somamos os coeficientes respectivos. Isso pressupõe que o efeito de cada uma das variáveis é independente. Contudo, isso nem sempre é o caso nos dados.

Vejamos um exemplo de Gries (2019, p.223). Imagine um estudo que compara a extensão de sujeitos e objetos em orações principais e em orações subordinadas, por meio do número de sílabas. O pesquisador coletou um *corpus*, separou sentenças com verbos transitivos diretos em que havia um SN sujeito e um SN objeto, contou o número de sílabas para cada SN, e codificou cada SN para as variáveis função sintática e tipo de oração.

A Figura 12.6 representa um resultado hipotético de um modelo linear que incluiu extensão em sílabas como variável resposta e função sintática e tipo de oração como variáveis previsoras. Os pontos representam as médias previstas do número de sílabas do SN para cada uma das condições.

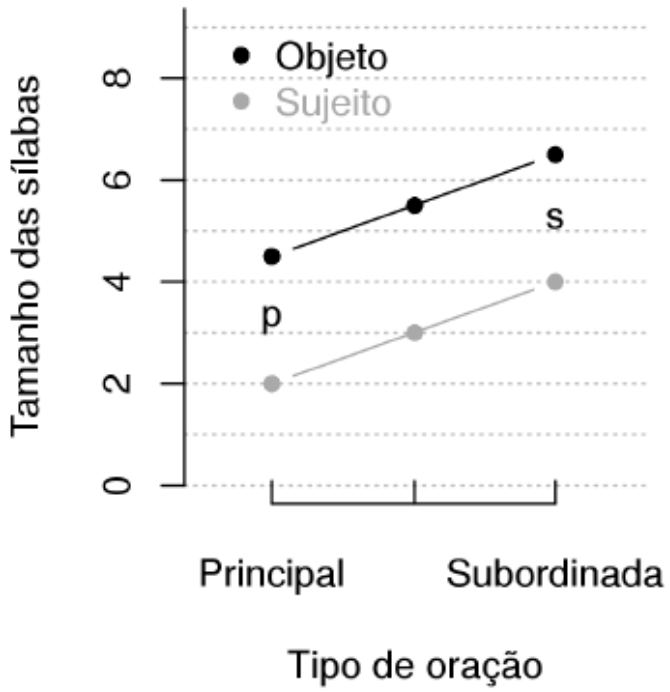

Figura 12.6: Exemplo de não interação entre VIs. Fonte: Gries (2019, p. 223).

A partir da figura, qual é a extensão prevista de sujeitos em orações principais?

- 2 sílabas
- 4 sílabas
- 4,5 sílabas
- 6,5 sílabas

A partir da figura, qual é a extensão prevista de objetos em orações subordinadas?

- 2 sílabas
- 4 sílabas
- 4,5 sílabas
- 6,5 sílabas

Nesse modelo, a extensão de SNs é em média maior em orações subordinadas do que em orações coordenadas, e em objetos em relação a sujeitos. A relação entre ambas as condições é constante: em média, as orações subordinadas têm SNs com 2 sílabas a mais do que em orações principais (tanto para sujeitos quanto para objetos), e objetos têm em média 2,5 sílabas a mais do que sujeitos (tanto em orações principais quanto em orações subordinadas). Os efeitos de ambas as variáveis são aditivos, de modo que se espera o menor número de sílabas em SNs sujeito em orações principais, e o maior número de sílabas em SNs objeto em orações subordinadas. A independência entre variáveis pode ser visualizada pelas linhas paralelas no gráfico.

A Figura 12.7 representa outro resultado hipotético a partir dos dados. O gráfico à esquerda mostra as estimativas de extensão média do SN em cada uma das quatro condições, e a figura à direita mostra, por meio da linha pontilhada, o que se esperaria caso a relação entre as variáveis fosse de independência.

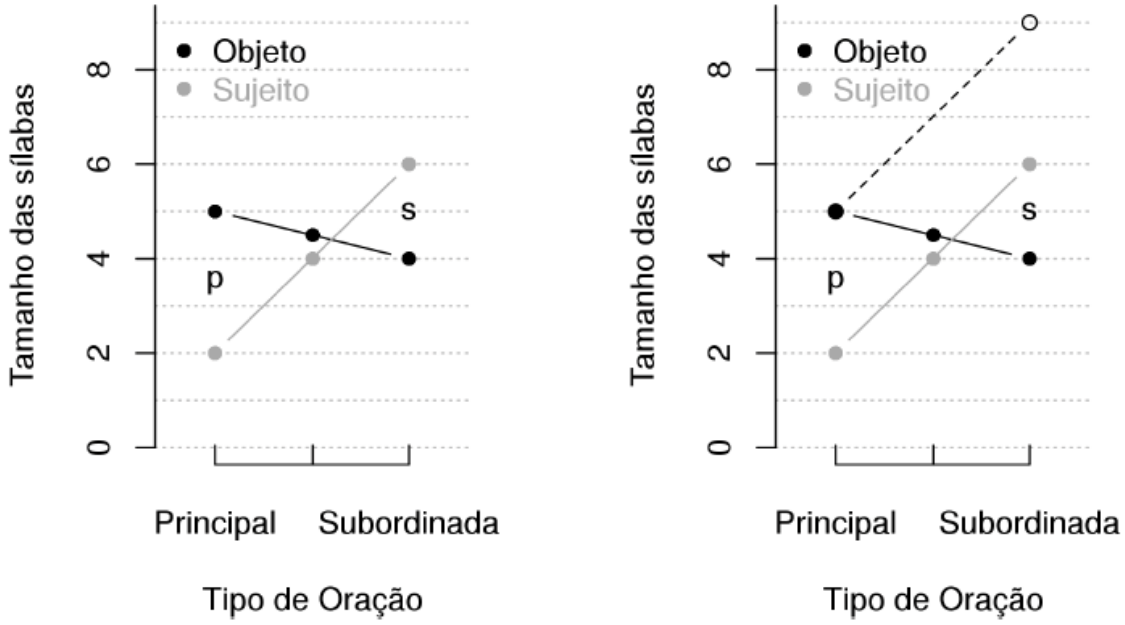

Figura 12.7: Exemplo de interação entre VIs (1). Fonte: Gries (2019, p. 223).

A partir da figura à esquerda, qual é a estimativa de extensão do SN sujeito em orações principais nesse modelo?

- 2 sílabas

- 4 sílabas
- 5 sílabas
- 6 sílabas

A partir da figura à esquerda, qual é a estimativa de extensão do SN objeto em orações subordinadas nesse modelo?

- 2 sílabas
- 4 sílabas
- 5 sílabas
- 6 sílabas

A figura à esquerda ilustra uma interação entre as variáveis função sintática e tipo de oração. A relação entre elas não é constante: enquanto para SNs sujeitos a diferença de extensão é de 4 sílabas entre orações principais e subordinadas, a diferença para SNs objetos não só é menor (1 sílaba), mas também vai na direção oposta (SNs maiores em orações principais do que em orações subordinadas). A interação é claramente visualizada pelo fato de que as linhas não são paralelas, mas se cruzam. Um modelo que previsse o efeito aditivo entre as variáveis previsoras – i.e. tamanho das sílabas ~ função sintática + tipo de oração – poderia chegar ao resultado do gráfico à direita, em que a estimativa da extensão de SNs objeto em orações subordinadas está bastante equivocada.

Veja agora a Figura 12.8, que ilustra outro caso hipotético de interação. De modo semelhante ao exemplo anterior, o gráfico à esquerda mostra as verdadeiras médias de extensão dos SNs e o gráfico à direita ilustra o resultado a que se chegaria caso não se previsse uma interação entre as variáveis previsoras.

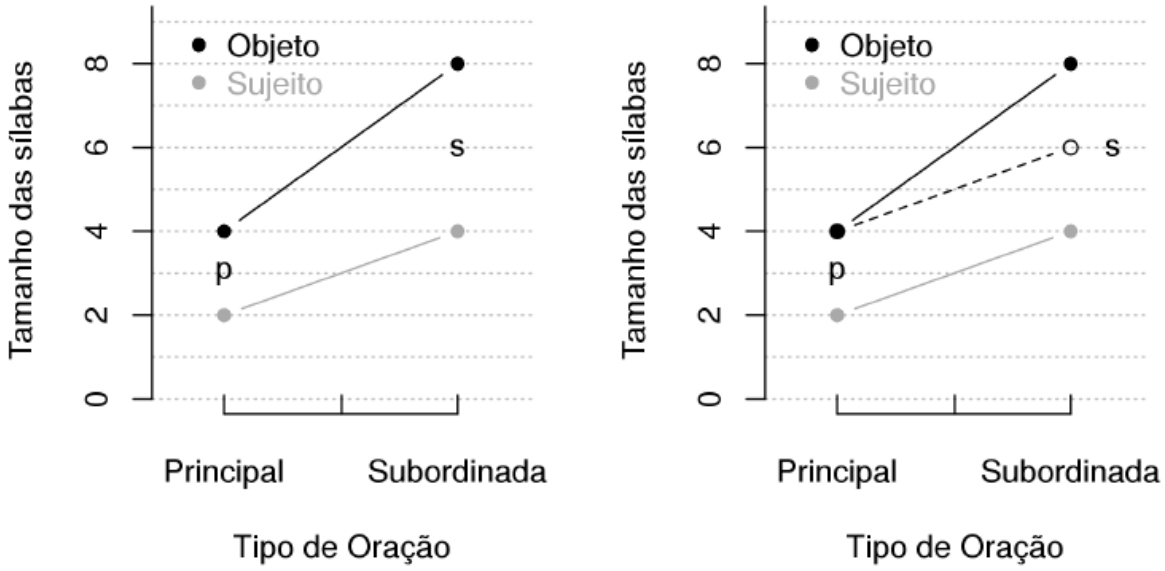

Figura 12.8: Exemplo de interação entre VIs (2). Fonte: Gries (2019, p. 225).

A partir da figura à esquerda, qual é a estimativa de extensão de SN sujeito em orações principais neste modelo?

- 2 sílabas
- 4 sílabas
- 6 sílabas
- 8 sílabas

A partir da figura à esquerda, qual é a estimativa de extensão de SN objeto em orações subordinadas neste modelo?

- 2 sílabas
- 4 sílabas
- 6 sílabas
- 8 sílabas

A partir da figura à direita, qual é a estimativa de extensão de SN objeto em orações subordinadas caso não se preveja a interação?

- 2 sílabas
- 4 sílabas

- 6 sílabas
- 8 sílabas

Aqui novamente se tem uma interação. Embora as linhas não se cruzem no gráfico, elas também não são paralelas. Isso significa que um modelo que não prevê a interação entre as variáveis função sintática e tipo de oração não seria capaz de estimar corretamente qual é a extensão dos SNs em cada condição. Para prever corretamente a extensão de SNs objeto em orações subordinadas, seria necessário ter mais um coeficiente que corrigisse a estimativa – neste caso, um coeficiente que informasse que ainda é necessário somar +2 para prever a extensão na condição objeto-oração subordinada.

Outro modo de entender o que é a interação entre duas variáveis previsoras é que o efeito de cada uma delas não pode ser determinado isoladamente, independentemente do efeito da outra. É necessário levar em consideração o efeito conjunto de ambas para bem prever a estimativa. Daí a importância de realizar análises multivariadas: seria impossível prever interações testando uma única variável a cada teste. Daí também deriva a inadequação de se usar o termo "variável independente" – em vez de "variável previsora" – quando se realiza uma análise multivariada, pois o efeito das variáveis não necessariamente é independente.

O ponto aqui é mostrar a importância de testar interações nos modelos lineares a fim de chegar a estimativas mais precisas para a variável resposta.

Você deve estar se perguntando como determinar se há interação entre duas variáveis. O R não tem um jeito automático de informar a você se duas variáveis previsoras são ou não são independentes entre si (nenhum software faz isso). Cabe ao pesquisador prever possíveis casos de interação, tentar visualizá-los em gráficos exploratórios e testá-los nos modelos estatísticos multivariados. A literatura sobre o assunto também é uma boa fonte para identificar casos de possível interação entre variáveis. Na Sociolinguística Variacionista, por exemplo, há muitos casos reportados de interação entre as variáveis classe social e sexo/gênero dos falantes (p.ex. Labov, 1990).

De posse dessa informação, um pesquisador que esteja trabalhando com essas duas variáveis previsoras já deve estar atento a um possível efeito interativo entre elas.

Nos modelos lineares, uma interação é incluída por meio do operador * em vez de +. No modelo `mod5`, vimos que há um efeito de AMOSTRA na estimativa de F1.NORM, mas não há um efeito de SEXO (não há diferença significativa entre homens e mulheres). A partir da linha de comando em que se criou `mod5`, crie um novo modelo `mod6` com a substituição de + por *. Nele, estamos testando não só o efeito das variáveis AMOSTRA e SEXO, como também se há uma interação entre elas.

```
mod6 <- lm(F1.NORM ~ AMOSTRA * SEXO, data = VOGAL_e2)
```

Visualize agora o resultado de `mod6` com `summary()`.

```
summary(mod6)

##
## Call:
## lm(formula = F1.NORM ~ AMOSTRA * SEXO, data = VOGAL_e2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.555 -17.028  -1.527  16.718  74.440
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 431.4651     2.1207 203.451  < 2e-16 **
*
## AMOSTRASP2010                -8.9387     2.9889  -2.991  0.00289 **
## SEXOmasculino                -1.8616     2.8118  -0.662  0.50816
## AMOSTRASP2010:SEXOmasculino   0.9562     3.9534   0.242  0.80896
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.45 on 673 degrees of freedom
## Multiple R-squared:  0.02743,    Adjusted R-squared:  0.02309
## F-statistic: 6.327 on 3 and 673 DF,  p-value: 0.0003107
```

Olhando os coeficientes, vemos que existe uma diferença significativa entre as amostras, e que não há diferença entre os sexos (como já visto acima). O modelo agora inclui um novo coeficiente, da interação entre AMOSTRASP2010:SEXOmasculino, que estima o valor de 0,9562 Hz. Isso significa que a estimativa de F1.NORM para falantes paulistanos do sexo masculino é 431,4651 - 8,9387 - 1,8616 + 0,9562 – ou seja, além dos coeficientes para AMOSTRASP2010 e para SEXOmasculino, há um novo coeficiente "de

ajuste". Contudo, neste modelo, o coeficiente da interação, 0,9562, não difere significativamente de zero, de modo que não faz diferença somar ou não esse novo coeficiente. Tal valor já estava previsto no intervalo de confiança do modelo sem interação. Pode-se concluir que as variáveis AMOSTRA e SEXO são independentes (e que SEXO não se correlaciona com a altura da vogal /e/ pretônica).

Podemos visualizar essa falta de interação por meio das funções `plot()` e `allEffects()`. A partir da linha de comando em que você usou essas duas funções, substitua o nome do modelo para `mod6`.

```
plot(allEffects(mod6), ylim = c(450, 400), grid = T)
```

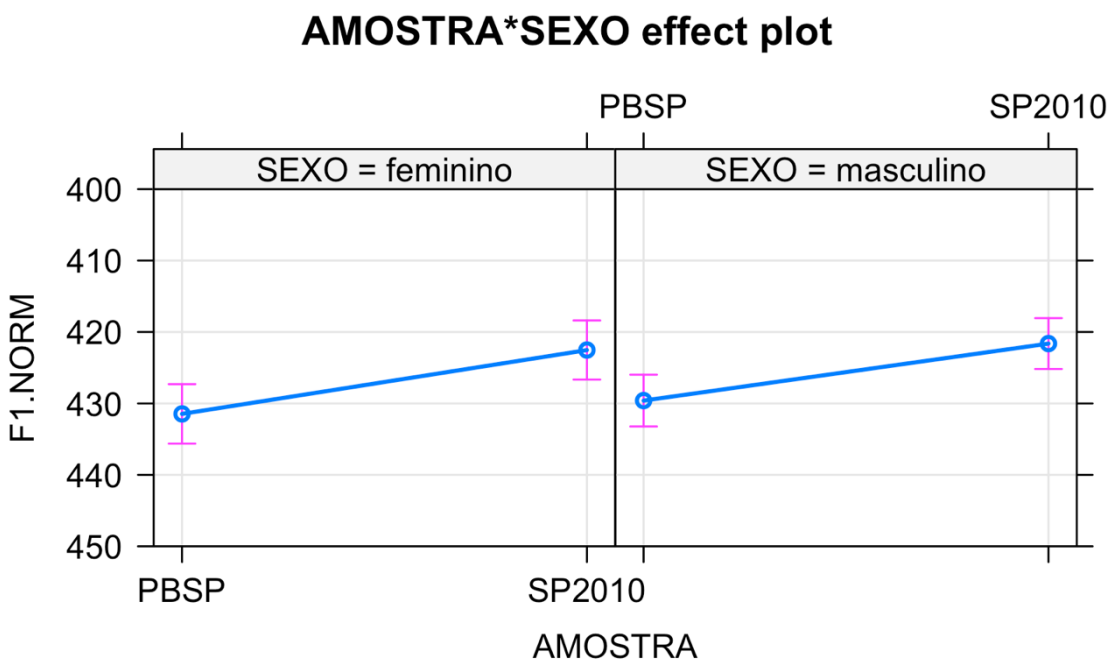

Figura 12.9: Gráfico de efeitos da interação entre as variáveis Amostra e Sexo para a altura da vogal /e/ pretônica. Fonte: própria.

Para o modelo `mod5`, em que havíamos incluído AMOSTRA + SEXO, foram plotados dois gráficos de efeitos, um para cada variável. Para o modelo `mod6`, em que incluímos AMOSTRA * SEXO, os gráficos incluem ambas as variáveis em cada painel. Essa figura é semelhante àquelas dos exemplos de Gries (basta imaginar ambos os painéis no mesmo plano). Aí vemos retas paralelas, que indicam independência entre as variáveis; a

diferença entre as amostras PBSP e SP2010 é significativa e constante tanto para falantes do sexo feminino quanto do sexo masculino.

Isso conclui a primeira parte sobre modelos de regressão linear, mas não conclui os passos da análise! Na próxima lição, veremos como avaliar a inclusão ou não de novas variáveis no modelo e como checar se nosso modelo não viola os pressupostos de um modelo linear.

Para saber mais

Recomendo fortemente a leitura dos capítulos 7 e 8 de Levshina (2015) para se aprofundar nos preceitos da análise de regressão linear. Esses capítulos apresentam os passos e os pressupostos de modelos lineares de modo bastante detalhado.

Exercícios

Nesta lista de exercícios, você vai desenvolver uma análise semelhante à que fizemos na Lição 12, mas agora sobre a vogal /o/ pretônica. Primeiro, carregue os dados da planilha Pretonicas.csv e crie um subconjunto de dados da vogal /o/ pretônica.

Nos dados de /e/ pretônica, a variável CONT.SEG foi recodificada de modo a juntar os segmentos consonantais de acordo com ponto/modo de articulação – ver *script* da Lição 12. Recodifique os dados das variáveis CONT.PREC e CONT.SEG do conjunto de dados da vogal /o/ pretônica com os mesmos critérios empregados para a vogal /e/ pretônica.

1. Há quantos dados de vogal /o/ pretônica na planilha Pretonicas.csv?
2. Há quantos dados de vibrantes no contexto seguinte à vogal pretônica /o/?
3. Há quantos dados de consoantes labiais no contexto precedente à vogal pretônica /o/?
4. Crie um modelo linear para testar se há correlação entre a altura da vogal pretônica (F1.NORM) e a origem do falante (AMOSTRA). Há diferença significativa entre paraibanos e paulistanos quanto à sua realização da vogal pretônica /o/? Justifique sua resposta.

5. Neste modelo, a distribuição dos resíduos segue a distribuição normal? Justifique sua resposta.
6. Faça um boxplot da distribuição de F1.NORM por AMOSTRA. Se existem valores atípicos, parece adequado excluir dados cujas medidas de F1.NORM estão acima de qual ponto?
    a. 600 Hz
    b. 550 Hz
    c. 450 Hz
    d. não há valores atípicos
7. Faça um novo subconjunto de dados incluindo apenas aqueles que estão abaixo do limite estipulado na questão anterior. Quantos dados foram efetivamente excluídos?
8. Refaça a análise para verificar se há correlação entre F1.NORM e AMOSTRA. A nova distribuição de resíduos segue a distribuição normal? Justifique sua resposta.
9. Faça um gráfico de efeitos do modelo acima. Há sobreposição entre os níveis dos intervalos de confiança da variável AMOSTRA? Explique sua resposta.
10. Crie um modelo que testa se há correlação entre a altura da vogal /o/ pretônica (F1.NORM) e o contexto precedente à vogal (CONT.PREC). Entre quais níveis da variável CONT.PREC há diferença significativa?
    a. entre labial e dental.alveolar
    b. entre velar e palatal.sibilante
    c. entre vibrante e labial
    d. entre palatal.sibilante e labial
11. Em quantos Hz a estimativa de F1.NORM para vibrante difere da estimativa do nível de referência?
12. Qual é a medida média de F1.NORM para a vogal /o/ quando é precedida por uma consoante vibrante?
13. O quanto da variação em F1.NORM é explicado pela variável CONT.PREC?

a. 0,8%
b. 1,5%
c. 3,4%
d. 3,9%
e. 27,5%

14. Teste se há interação entre as variáveis AMOSTRA e CONT.PREC. Neste modelo, há interação entre AMOSTRA e CONT.PREC? Justifique sua resposta.
15. Qual é a medida média de F1.NORM para a vogal /o/ quando precedida de consoante velar na fala de paraibanos?
16. Qual é a medida média de F1.NORM para a vogal /o/ quando precedida de consoante labial na fala de paulistanos?
17. Crie um modelo para testar se há correlação entre a altura da vogal /o/ pretônica (F1.NORM) e a altura da vogal da sílaba seguinte (F1.SEG.NORM). A cada unidade de F1.SEG.NORM, em quanto se modifica a estimativa de F1.NORM?
    a. 0,03006
    b. 0,0689
    c. 0,20859
    d. 6,938
18. Calcule a estimativa da medida de F1.NORM quando F1.SEG.NORM tem 450 Hz.
19. Vimos três testes estatísticos que podem ser aplicados a VDs numéricas: (i) teste-t, (ii) teste de correlação e (iii) regressão linear. Às vezes podemos aplicar mais de um deles, às vezes não. Se o pesquisador quer testar se há correlação entre uma VD numérica e uma VI numérica, qual(is) teste(s) pode(m) ser aplicado(s)?
    a. apenas (i)
    b. apenas (iii)
    c. (i) e (ii)
    d. (ii) e (iii)

20. Considere (i) teste-t, (ii) teste de correlação e (iii) regressão linear. Se um pesquisador quer testar se há correlação entre uma VD numérica e uma VI nominal binária, qual(is) teste(s) pode(m) ser aplicado(s)?
    a. apenas (i)
    b. apenas (iii)
    c. (i) e (iii)
    d. (ii) e (iii)
21. Considere (i) teste-t, (ii) teste de correlação e (iii) regressão linear. Se um pesquisador quer testar se há correlação entre uma VD numérica, uma VI nominal com 5 fatores e uma VI numérica, qual(is) teste(s) pode(m) ser aplicado(s)?
    a. apenas (ii)
    b. apenas (iii)
    c. (i) e (iii)
    d. (ii) e (iii)