Lição 14: Regressão Logística 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 DadosRT.csv.

```
# Definir diretório de trabalho

#setwd()

# Importar dados da planilha

dados <- read_csv("DadosRT.csv",
                  col_types = cols(.default = col_factor(),
                                   VD = col_factor(levels = c("tepe",
"retroflexo")),
                                   FAIXA.ETARIA = col_factor(levels =
c("1a", "2a", "3a")),
                                   ESCOLARIDADE = col_factor(levels =
c("fundamental", "medio", "superior")),
                                   REGIAO = col_factor(levels = c("cen
tral", "periferica")),
                                   CONT.FON.PREC = col_factor(levels =
c("i", "e", "3", "a", "0", "o", "u")),
                                   TONICIDADE = col_factor(levels = c(
"atona", "tonica")),
                                   POSICAO.R = col_factor(levels = c("
final", "medial")),
                                   CLASSE.MORFOLOGICA = col_factor(lev
els = c("adjetivo", "adverbio", "conj.prep", "morf.inf", "substantivo"
, "verbo")),
                                   IDADE = col_integer(),
                                   INDICE.SOCIO = col_double(),
                                   FREQUENCIA = col_double()
                                   )
                  )

dados$CONT.FON.SEG <- fct_collapse(dados$CONT.FON.SEG,
                                   pausa = "#",
                                   coronal = c("t", "d", "s", "z", "x"
, "j", "ts", "dz", "l", "n"),
                                   labial = c("p", "b", "f", "v", "m")
,
                                   dorsal = c("k", "g", "h")
                                   )

dados$CONT.FON.SEG <- fct_relevel(dados$CONT.FON.SEG, "pausa", "corona
l", "dorsal", "labial")

###Funções úteis (Gries 2019)
```

```
logit <- function(x) {
  log(x/(1-x))
  }

ilogit <- function(x) {
  1/(1+exp(-x))
  }
```

Esta e a próxima lição são dedicadas a análises de regressão logística, que se aplicam a variáveis dependentes/resposta binárias. A regressão logística permite a inclusão de múltiplas variáveis previsoras, assim como a análise de regressão linear. Cabe lembrar que, antes de chegar à análise multivariada de regressão logística, o pesquisador idealmente já terá feito gráficos e análises exploratórias (como o qui-quadrado) a fim de saber como se distribuem seus dados. O interesse nas análises de regressão logística é verificar o efeito simultâneo de múltiplas variáveis previsoras, a fim de chegar a um modelo para descrever, explicar e prever o comportamento da variável resposta.

A regressão logística também gera um coeficiente linear (Intercept) e coeficientes angulares para cada variável/termo previsor do modelo, e avalia se a estimativa difere significativamente de zero. Contudo, enquanto a regressão linear gera coeficientes na mesma unidade que da variável resposta – nos exemplos das Lições 12 e 13, medidas de F1.NORM em Hz –, a regressão logística gera coeficientes em *logodds* (também chamado de log-odds-ratio), sobre o qual falaremos mais adiante. Os valores de logodds entram na função y = a + bx1, cx2, dx3... para permitir a estimativa de probabilidade de ocorrência dos níveis da variável resposta.

Carregue inicialmente os pacotes necessários para esta lição: tidyverse, car, effects e rms.

```
library(tidyverse)
library(car)
library(effects)
library(rms)
```

Deixei disponível para você o conjunto de dados da pronúncia do /r/ em coda silábica como tepe ou como retroflexo na fala de paulistanos, num dataframe chamado `dados`. Aplique a função `str()` para se (re)familiarizar com ele.

```
str(dados)

## spec_tbl_df [9,226 × 20] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ VD                : Factor w/ 2 levels “tepe”,”retroflexo”: 2 2
2 1 2 2 2 2 1 2 ...
##  $ PARTICIPANTE      : Factor w/ 118 levels “IvanaB”,”HeloisaS”,..:
1 1 1 1 1 1 1 1 1 1 ...
##  $ SEXO.GENERO       : Factor w/ 2 levels “feminino”,”masculino”: 1
1 1 1 1 1 1 1 1 1 ...
##  $ IDADE             : int [1:9226] 30 30 30 30 30 30 30 30 30 30 .
..
##  $ FAIXA.ETARIA      : Factor w/ 3 levels “1a”,”2a”,”3a”: 1 1 1 1 1
1 1 1 1 1 ...
##  $ ESCOLARIDADE      : Factor w/ 3 levels “fundamental”,..: 1 1 1 1
1 1 1 1 1 1 ...
##  $ REGIAO            : Factor w/ 2 levels “central”,”periferica”: 2
2 2 2 2 2 2 2 2 ...
##  $ INDICE.SOCIO      : num [1:9226] 2 2 2 2 2 2 2 2 2 2 ...
##  $ ORIGEM.PAIS       : Factor w/ 5 levels “mista”,”SPcapital”,..: 1
1 1 1 1 1 1 1 1 ...
##  $ CONT.FON.PREC     : Factor w/ 7 levels “i”,”e”,”3”,”a”,..: 4 6 2
2 4 4 5 4 5 3 ...
##  $ CONT.FON.SEG      : Factor w/ 4 levels “pausa”,”coronal”,..: 2 2
3 4 3 2 2 2 3 2 ...
##  $ TONICIDADE        : Factor w/ 2 levels “atona”,”tonica”: 2 1 1 1
2 2 2 2 1 2 ...
##  $ POSICAO.R         : Factor w/ 2 levels “final”,”medial”: 2 2 2 2
1 1 2 2 2 2 ...
##  $ CLASSE.MORFOLOGICA: Factor w/ 6 levels “adjetivo”,”adverbio”,..:
5 5 5 5 5 5 5 5 3 5 ...
##  $ FREQUENCIA        : num [1:9226] 1.34 0.16 0.22 0.44 1.94 1.94 0
.35 0.03 5.98 0.16 ...
##  $ ESTILO            : Factor w/ 4 levels “conversacao”,..: 1 1 1 1
1 1 1 1 1 1 ...
##  $ ITEM.LEXICAL      : Factor w/ 1151 levels “parte”,”jornal”,..: 1
2 3 4 5 5 6 7 8 9 ...
##  $ cont.precedente   : Factor w/ 6836 levels “do CEU é daquela”,..:
1 2 3 4 5 6 7 8 9 10 ...
##  $ ocorrencia        : Factor w/ 1760 levels “parte <R>“,”jornal <R
>“,..: 1 2 3 4 5 5 6 7 8 9 ...
##  $ cont.seguinte     : Factor w/ 6813 levels “que as perua(s) ia”,.
.: 1 2 3 4 5 6 7 8 9 10 ...
##  - attr(*, “spec”)=
##   .. cols(
##   ..   .default = col_factor(),
##   ..   VD = col_factor(levels = c(“tepe”, “retroflexo”), ordered =
FALSE, include_na = FALSE),
##   ..   PARTICIPANTE = col_factor(levels = NULL, ordered = FALSE, in
clude_na = FALSE),
##   ..   SEXO.GENERO = col_factor(levels = NULL, ordered = FALSE, inc
lude_na = FALSE),
##   ..   IDADE = col_integer(),
##   ..   FAIXA.ETARIA = col_factor(levels = c(“1a”, “2a”, “3a”), orde
red = FALSE, include_na = FALSE),
```

```
##   ..   ESCOLARIDADE = col_factor(levels = NULL, ordered = FALSE, in
clude_na = FALSE),
##   ..   REGIAO = col_factor(levels = c("central", "periferica"), ord
ered = FALSE, include_na = FALSE),
##   ..   INDICE.SOCIO = col_double(),
##   ..   ORIGEM.PAIS = col_factor(levels = NULL, ordered = FALSE, inc
lude_na = FALSE),
##   ..   CONT.FON.PREC = col_factor(levels = c("i", "e", "3", "a", "0
", "o", "u"), ordered = FALSE, include_na = FALSE),
##   ..   CONT.FON.SEG = col_factor(levels = NULL, ordered = FALSE, in
clude_na = FALSE),
##   ..   TONICIDADE = col_factor(levels = c("atona", "tonica"), order
ed = FALSE, include_na = FALSE),
##   ..   POSICAO.R = col_factor(levels = c("final", "medial"), ordere
d = FALSE, include_na = FALSE),
##   ..   CLASSE.MORFOLOGICA = col_factor(levels = c("adjetivo", "adve
rbio", "conj.prep", "morf.inf", "substantivo",
##   ..     "verbo"), ordered = FALSE, include_na = FALSE),
##   ..   FREQUENCIA = col_double(),
##   ..   ESTILO = col_factor(levels = NULL, ordered = FALSE, include_
na = FALSE),
##   ..   ITEM.LEXICAL = col_factor(levels = NULL, ordered = FALSE, in
clude_na = FALSE),
##   ..   cont.precedente = col_factor(levels = NULL, ordered = FALSE,
include_na = FALSE),
##   ..   ocorrencia = col_factor(levels = NULL, ordered = FALSE, incl
ude_na = FALSE),
##   ..   cont.seguinte = col_factor(levels = NULL, ordered = FALSE, i
nclude_na = FALSE)
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Aplique também a função `View()` para visualizar a planilha. Em especial, veja as variáveis SEXO.GENERO, FAIXA.ETARIA, INDICE.SOCIO e REGIAO, com que trabalharemos nesta lição.

```
View(dados)
```

N.B.: Resultado aqui omitido.

A variável SEXO.GENERO indica se o dado veio de um falante do sexo feminino ou masculino. FAIXA.ETARIA está dividida em três níveis – 1a: de 20 a 34 anos; 2a: de 35 a 59 anos; e 3a: 60 anos ou mais. REGIAO indica a região de residência atual do falante, central ou periférica. Por fim, INDICE.SOCIO é um índice contínuo que vai de 1 a 5, que leva em conta a escolaridade, ocupação e renda do falante, bem como escolaridade e ocupação de seus pais (ver Oushiro, 2015, cap.3 para detalhes); quanto maior o índice, maior foi a pontuação média do falante para os critérios acima.

De modo semelhante ao que fizemos na análise de regressão linear, vamos começar com modelos simples, com o intuito de treinar a leitura dos resultados. Na próxima lição, veremos como criar e avaliar modelos mais complexos, que são aqueles que você efetivamente vai querer reportar nas publicações.

A função para criar um modelo de regressão logística é `glm()` – do inglês, "generalized linear model". Trata-se efetivamente de uma generalização do modelo de regressão linear para variáveis não numéricas. Você verá que muito da implementação da regressão logística é paralela à regressão linear. Mas é importante já mencionar uma diferença importante: as estimativas da regressão logística são fornecidas em relação ao *segundo* nível da variável dependente. Aplique a função `levels()` ao vetor `dados$VD` para ver os níveis da variável dependente.

```
levels(dados$VD)
## [1] "tepe"       "retroflexo"
```

O R organizou os níveis na ordem tepe e retroflexo a partir da especificação de levels no momento da importação dos dados (ver *script*). Isso significa que os resultados dos modelos deverão ser lidos em termos do que aumenta ou diminui a probabilidade de ocorrência do *retroflexo*. Essa variante foi escolhida para leitura dos resultados pois temos mais interesse em vê-los da perspectiva da variante não prototípica da comunidade paulistana. Será mais interessante descobrir quais fatores mais favorecem o emprego de retroflexo do que do tepe.

Podemos então seguir para a regressão logística. Assim como a função `lm()`, o primeiro argumento de `glm()` é uma fórmula no formato VR ~ VP..., e o segundo argumento é o conjunto de dados. Aqui, entretanto, há mais um argumento: family = binomial, que indica que a variável resposta é binária. Crie então um primeiro modelo chamado `mod1`, que testa se há correlação entre a variação na pronúncia de /r/ em coda (VD) e o SEXO.GENERO do falante. Digite `mod1 <- glm(VD ~ SEXO.GENERO, data = dados, family = binomial)`.

```
mod1 <- glm(VD ~ SEXO.GENERO, data = dados, family = binomial)
```

Veja o resultado com a aplicação de `summary()` a `mod1`.

```
summary(mod1)
```

```
##
## Call:
## glm(formula = VD ~ SEXO.GENERO, family = binomial, data = dados)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8899  -0.8899  -0.7375   1.4952   1.6941
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -1.16304    0.03475 -33.470   <2e-16 ***
## SEXO.GENEROmasculino  0.44111    0.04672   9.442   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10993  on 9225  degrees of freedom
## Residual deviance: 10903  on 9224  degrees of freedom
## AIC: 10907
##
## Number of Fisher Scoring iterations: 4
```

Você já está bem familiarizado com o *output* de um teste estatístico no R. Primeiro, ele informa a fórmula empregada e os resíduos. Em seguida, mostra a tabela de coeficientes, com o erro padrão e um valor de significância para cada um. Por fim, dá informações gerais sobre o modelo como um todo. Mas há aqui algumas pequenas diferenças.

A principal delas está nas estimativas dos coeficientes, que são dadas em logodds. Quando fizemos um modelo de regressão linear, era fácil interpretar as estimativas, pois elas eram dadas na mesma unidade que a variável resposta. Mas, aqui, o que significam -1,16304 logodds para o coeficiente linear e 0,44111 logodds para o coeficiente angular? Na sequência, vamos ver como esse valor é calculado para você mais bem entendê-lo.

Visualize as frequências dos dados de SEXO.GENERO pela VD, substituindo no *script* os termos necessários.

```
dados %>%
  count(SEXO.GENERO, VD)
```

```
## # A tibble: 4 × 3
##   SEXO.GENERO VD            n
##   <fct>       <fct>     <int>
## 1 feminino    tepe       3478
```

```
## 2 feminino    retroflexo  1087
## 3 masculino   tepe        3137
## 4 masculino   retroflexo  1524
```

As chances (= odds) de algo ocorrer se dão pela divisão simples entre o número de resultados favoráveis e o número de resultados desfavoráveis. Embora relacionadas com a probabilidade, as chances são algo diferente. As chances normalmente são expressas por uma razão, do tipo 3:1. Assim, as chances de ocorrer *tepe* na fala de mulheres são 3478 / 1087, de acordo com a tabela acima. Guarde esse resultado num objeto chamado `odds_F`.

```
odds_F <- 3478 / 1087
```

Faça o mesmo cálculo para os homens, e guarde-o num objeto chamado `odds_M`.

```
odds_M <- 3137 / 1524
```

Faça agora o cálculo do odds-ratio, a razão (= divisão) entre as chances de ocorrer tepe na fala das mulheres (odds_F) e na fala dos homens (odds_M). Guarde-o num objeto chamado `odds_ratio`.

```
odds_ratio <- odds_F / odds_M
```

Agora aplique a função `log()` a `odds_ratio`.

```
log(odds_ratio)
```

```
## [1] 0.4411073
```

Veja que o resultado do log de odds-ratio é justamente a estimativa calculada para os homens na tabela de coeficientes de nosso modelo – então agora você sabe de onde saiu essa estimativa. Por que usar o log do odds-ratio, e não simplesmente as chances (= odds) ou a probabilidade (= proporção de ocorrências da variante de interesse em relação ao total)?

A Figura 14.1 (Gries, 2019, p. 265) mostra a relação entre essas três medidas de probabilidade. Odds é uma escala de vai de zero até +∞. Isso não é difícil de entender. As chances de algo ocorrer (=resultado favorável – Fav) são maiores, menores ou iguais às chances de não ocorrer (=resultado desfavorável – Des). Se o resultado favorável é mais frequente do que o resultado desfavorável, a divisão Fav/Des dará um número maior do que 1, e não tem limite máximo. Se o resultado favorável é menos frequente do que o

resultado desfavorável, a divisão Fav/Des dará um número menor do que 1, mas nunca negativo. O ponto neutro da escala é 1, pois ele equivale ao cenário em que Fav = Des, portanto Fav/Des = 1.

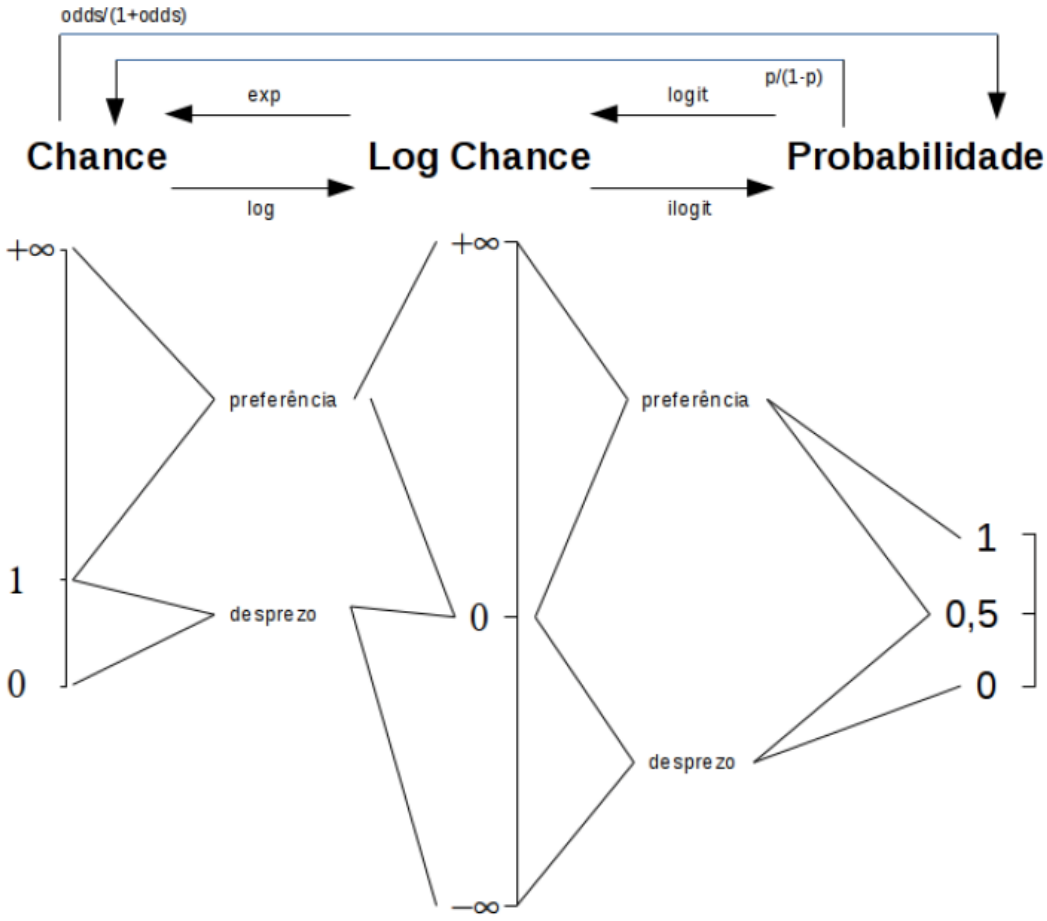

Figura 14.1: Relações entre as medidas de odds, logodds e probabilidade. Fonte: Gries (2019, p. 265).

Por outro lado, a probabilidade é uma medida mais conhecida. Ela é uma escala de 0 a 1, em que 0 representa uma chance nula de algo ocorrer (0%) e 1 representa certeza de que vai ocorrer (100%). De modo simples, ela é calculada pelo número de resultados favoráveis pelo total de observações (Fav/T). O ponto neutro aí é 50%; um número abaixo disso indica maior probabilidade de que o evento não vai ocorrer e um número acima disso indica maior probabilidade de que o evento vai ocorrer.

Por fim, a escala de logodds vai de menos infinito a mais infinito, com ponto neutro em zero. Em relação a odds, ela tem a vantagem de ser uma escala simétrica, com

um mesmo intervalo entre o ponto neutro e suas extremidades. A operação log tem justamente o papel de transformar valores entre 0 e 1 em um valor negativo (experimente depois no Console do R!). Desse modo, a interpretação de valores é muito mais intuitiva do que numa escala assimétrica, pois os intervalos em que há favorecimento ou desfavorecimento de um evento são diretamente comparáveis.

Em relação à escala de probabilidade, ter o ponto neutro em zero – em vez de 0,5 – também traz vantagens. Vimos, no modelo de regressão linear, que o resultado em termos de *diferença* em relação ao intercept permite avaliar mais prontamente se ela é zero ou não. Um logodds de zero (ou próximo a ele) indica prontamente se há diferenças significativas; valores positivos indicam tendência a favorecimento (em relação a outro nível da mesma variável previsora); e valores negativos indicam tendência a desfavorecimento (em relação a outro nível da mesma variável previsora). No resultado de nosso primeiro modelo, portanto, um logodds de 0,44 para homens indica que, *em relação às mulheres*, os homens favorecem o retroflexo.

Coloquei a comparação entre homens e mulheres acima em destaque porque uma estimativa de logodds positiva não significa que os homens usam o retroflexo mais frequentemente do que o tepe. Essa é uma diferença relativa. Lembre-se que para chegar à estimativa, é necessário somar o valor do intercept ao valor do coeficiente angular. Os homens paulistanos, portanto, têm -1.16304 + 0.44111 = -0.7219327 logodds de emprego de retroflexo. A Figura 14.1 mostra qual operação se aplica para transformar uma medida em outra – exp, log, logit ou ilogit. Aplique a função `ilogit()` a -0.7219327 para transformar a medida de logodds em probabilidade.

```
ilogit(-0.7219327)
```

```
## [1] 0.3269675
```

O cálculo acima indica que a probabilidade de os homens empregarem retroflexo é 32,7%. Aplique também a função `ilogit()` ao valor de -1.16304, a estimativa em logodds para as mulheres.

```
ilogit(-1.16304)
```

```
## [1] 0.2381153
```

O cálculo para as mulheres indica uma probabilidade de 23,8% de elas empregarem o retroflexo. Para comparar, faça agora uma tabela de proporções, substituindo os termos df e VAR na estrutura pré-montada do *script* – reveja a Lição 4, se necessário.

```
dados %>%
  count(SEXO.GENERO, VD) %>%
  group_by(SEXO.GENERO) %>%
  mutate(prop = prop.table(n))
```

```
## # A tibble: 4 × 4
## # Groups:   SEXO.GENERO [2]
##   SEXO.GENERO VD             n  prop
##   <fct>       <fct>      <int> <dbl>
## 1 feminino    tepe        3478 0.762
## 2 feminino    retroflexo  1087 0.238
## 3 masculino   tepe        3137 0.673
## 4 masculino   retroflexo  1524 0.327
```

O cálculo de probabilidade, nesse modelo univariado, corresponde exatamente ao valor de logodds, pois não há outras variáveis previsoras que possam ajustar a estimativa de logodds. Legal, não?

Não é necessário memorizar toda essa demonstração, nem a realizar todas as vezes que criar um modelo de regressão logística. Ela serve aqui para que você perceba que as medidas de logodds se relacionam diretamente a outras medidas estatísticas mais conhecidas (chances e probabilidades), mas têm a vantagem de ser uma escala centrada em zero, o que facilita a interpretação dos resultados.

As medidas em logodds podem ser estranhas no começo, pela falta de contato que se tem com elas. No entanto, aos poucos você vai se acostumar a ver uma medida em logodds e saber se se trata de um efeito forte ou fraco. A Tabela 14.1 (adaptada de Levshina, 2015, p. 265) mostra as equivalências entre probabilidades, *odds* e *logodds*.

Tabela 14.1: Tabela de conversão de valores entre medidas de probabilidades, odds e logodds.

Probabilidade	*Odds*	Logit (*log odds*)
0,001	0,001	-6,91
0,01	0,01	-4,60
0,05	0,05	-2,94

0,10	0,11	-2,20
0,25	0,33	-1,10
0,50	1	0
0,75	3	1,10
0,90	9	2,20
0,95	19	2,94
0,99	99	4,60
0,999	999	6,91

Fonte: Levshina (2015, p. 265).

A essa altura, o resultado de `mod1` ficou lá atrás, mas ainda falta comentar as demais medidas estatísticas geradas no modelo. Digite então `summary(mod1)` no Console.

```
summary(mod1)

##
## Call:
## glm(formula = VD ~ SEXO.GENERO, family = binomial, data = dados)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8899  -0.8899  -0.7375   1.4952   1.6941
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -1.16304    0.03475 -33.470   <2e-16 ***
## SEXO.GENEROmasculino  0.44111    0.04672   9.442   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10993  on 9225  degrees of freedom
## Residual deviance: 10903  on 9224  degrees of freedom
## AIC: 10907
##
## Number of Fisher Scoring iterations: 4
```

De modo semelhante ao modelo linear, a regressão logística apresenta o erro padrão junto à estimativa. Se se faz a divisão entre Estimativa / Erro Padrão, chega-se ao valor da terceira coluna. Aqui, em vez de um valor-*t*, temos um valor-*z*, que é consultado numa tabela de distribuição normal padrão (como fizemos na tabela de qui-quadrado, na Lição 9). Dessa tabela se obtém o valor de significância.

Ao pé da tabela, diferentemente do modelo linear, o R não reporta diretamente medidas que permitem avaliar o poder explanatório do modelo, como R^2 e a estatística-F. No entanto, os valores de desvio nulo e desvio residual permitem fazer esse cálculo. O desvio nulo se refere a quanto há de variabilidade total nos dados, sem a inclusão de qualquer variável previsora. O desvio residual se refere a quanto há de variabilidade nos dados depois da inclusão da(s) variável(is) previsora(s). Portanto, a diferença entre o desvio residual e o desvio nulo é o quanto o nosso modelo é capaz de prever da variabililidade dos dados.

No Anexo C, deixei disponível um *script* que permite fazer o cálculo de significância do modelo como um todo "à mão", de posse dos dados de desvio nulo, desvio residual e graus de liberdade. Depois dessa lição, dedique um tempo para entendê-lo. Ele não será explicado aqui porque há uma maneira mais simples de se obter essa medida estatística, que veremos logo adiante.

O AIC (Akaike Information Criterion), como vimos na Lição 13, é uma medida que permite comparar modelos e é usada na função `step()` para selecionar ou excluir variáveis. E a última medida estatística reportada para o modelo logístico é o Fisher Scoring iterations. Ela se refere ao número de iterações do modelo até que os resultados convergiram. Aqui, o modelo convergiu após 4 tentativas. Quando este número é grande (digamos, acima de 20), é sinal de que foram incluídas mais variáveis previsoras do que é possível explicar com a quantidade de dados de que se dispõe. A solução, neste caso, é diminuir o número de variáveis incluídas no modelo.

De modo semelhante ao modelo linear, podemos plotar um gráfico de efeitos para mais bem visualizar os resultados numéricos. Para isso, vamos usar uma função do pacote effects, já carregado acima.

Para modelos logísticos, usamos a função `allEffects()` – mesmo se o modelo contém apenas uma variável previsora – e o argumento type = "response". Faça o gráfico de efeitos com `plot(allEffects(mod1), type = "response")` (Figura 14.2).

```
plot(allEffects(mod1), type = "response")
```

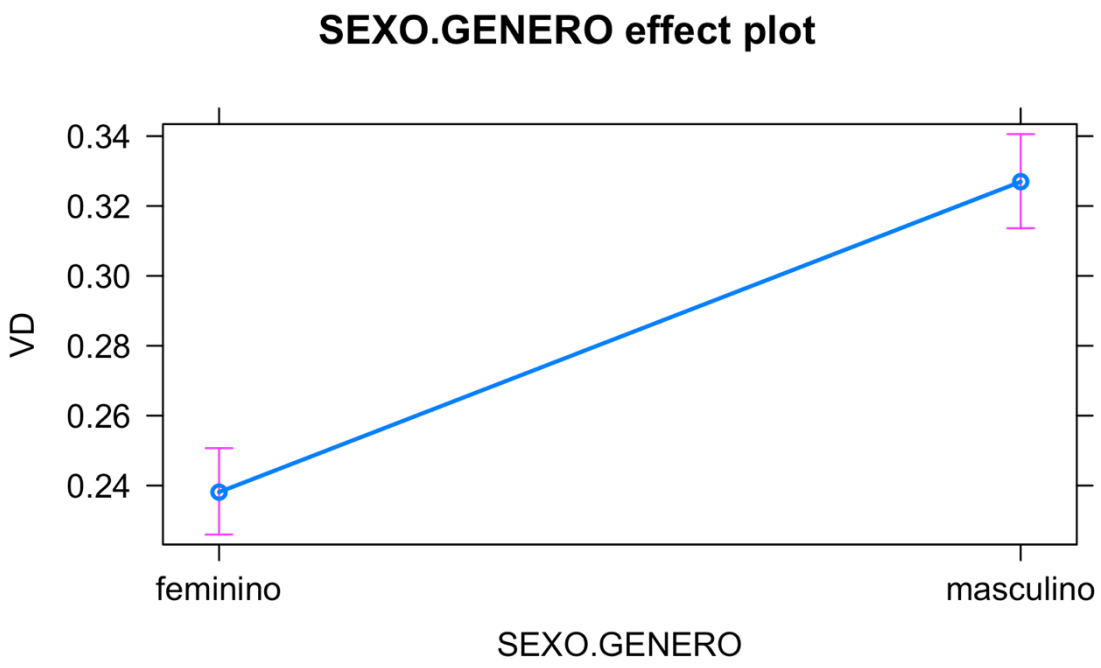

Figura 14.2: Gráfico de efeitos da variável Sexo para o uso de /r/ retroflexo. Fonte: própria.

O gráfico de efeitos mostra os resultados com as medidas de probabilidade. Vemos que a estimativa de emprego de retroflexo é cerca de 23% para as mulheres e 32% para os homens, e que os intervalos de confiança não se sobrepõem, o que permite inferir a diferença significativa entre os níveis.

Há outra função que faz regressões logísticas, chamada `lrm()` – "logistic regression model". Esta função pertence ao pacote rms, também carregado no início desta lição. A função `lrm()` toma como argumentos apenas a fórmula (VD ~ SEXO.GENERO) e o conjunto de dados (data = dados). Aplique-a então com `lrm(VD ~ SEXO.GENERO, data = dados)`.

```
lrm(VD ~ SEXO.GENERO, data = dados)

## Logistic Regression Model
##
##  lrm(formula = VD ~ SEXO.GENERO, data = dados)
##
##                  Model Likelihood    Discrimination    Rank Discrim
.
##                        Ratio Test           Indexes          Indexe
s
##Obs    9226    LR chi2      90.07    R2       0.014      C        0.55
```

```
5
##tepe  6615    d.f.             1    g        0.221      Dxy     0.10
9
##retr  2611    Pr(> chi2) <0.0001    gr       1.247      gamma   0.21
7
##max |deriv| 1e-11                   gp       0.044      tau-a   0.04
4
##                                    Brier    0.201
##
##                         Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept              -1.1630 0.0347 -33.47 <0.0001
##  SEXO.GENERO=masculino   0.4411 0.0467   9.44 <0.0001
##
```

Trata-se de outro método para o mesmo fim. Veja que os coeficientes gerados correspondem exatamente àqueles do modelo prévio com a função `glm()`. Mas, além de não ser necessário usar a função `summary()` para visualizar o resultado, a função `lrm()` apresenta algumas medidas estatísticas relevantes no topo do *output*. A primeira coluna fornece o total de observações e de cada variante da variável resposta.

A segunda coluna, Model Likelihood Ratio Test, informa se o modelo como um todo é significativo. O teste de Razão de Verossimilhança compara dois modelos (com e sem variáveis previsoras) e com isso gera um valor-*p* para o modelo – como é feito no *script* do Anexo C. As duas colunas à direita mostram várias medidas estatísticas de qualidade do ajuste, ou seja, de quão bem o modelo é capaz de explicar a variação encontrada nos dados (como o R^2). Para modelos de regressão logística, a estatística mais reportada é o índice de Concordância C, a primeira medida da quarta coluna. De Hosmer & Lemeshow 2000 (*apud* Levshina, 2015, p. 259), temos C = 0,5: pouco poder de discriminação de resultado; 0,7 < C < 0,8: poder aceitável de discriminação de resultado; 0,8 < C < 0,9: poder excelente de discriminação de resultado; e C > 0,9: poder notório de discriminação de resultado. Em nosso modelo, C = 0,55 é um sinal de que ele ainda pode ser melhorado.

Vamos seguir agora para um modelo com uma variável previsora nominal com mais de dois níveis. Com a função `glm()`, crie o modelo `mod2` que testa se há correlação entre VD e a FAIXA.ETARIA do falante. Não se esqueça de incluir o conjunto de dados e family = binomial.

```
mod2 <- glm(VD ~ FAIXA.ETARIA, data = dados, family = binomial)
```

E veja o resultado com `summary()`.

```
summary(mod2)

##
## Call:
## glm(formula = VD ~ FAIXA.ETARIA, family = binomial, data = dados)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9570  -0.7757  -0.7061   1.4151   1.7386
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.54330    0.03755 -14.468   <2e-16 ***
## FAIXA.ETARIA2a -0.50356    0.05484  -9.183   <2e-16 ***
## FAIXA.ETARIA3a -0.71874    0.05833 -12.322   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10993  on 9225  degrees of freedom
## Residual deviance: 10824  on 9223  degrees of freedom
## AIC: 10830
##
## Number of Fisher Scoring iterations: 4
```

Vamos direto aos coeficientes. O resultado mostra que, em relação à primeira faixa etária (nível de referência), os falantes tanto de segunda quanto de terceira faixa etária tendem a *des*favorecer o retroflexo – que se infere pelas estimativas negativas em logodds. Em outras palavras, os falantes mais jovens tendem a empregar mais retroflexos do que os mais velhos, o que pode ser indício de uma mudança em progresso na comunidade paulistana.

A partir da última linha de comando em que você usou `plot()` e `allEffects()`, mude o nome do modelo para `mod2` (Figura 14.3).

```
plot(allEffects(mod2), type = “response”)
```

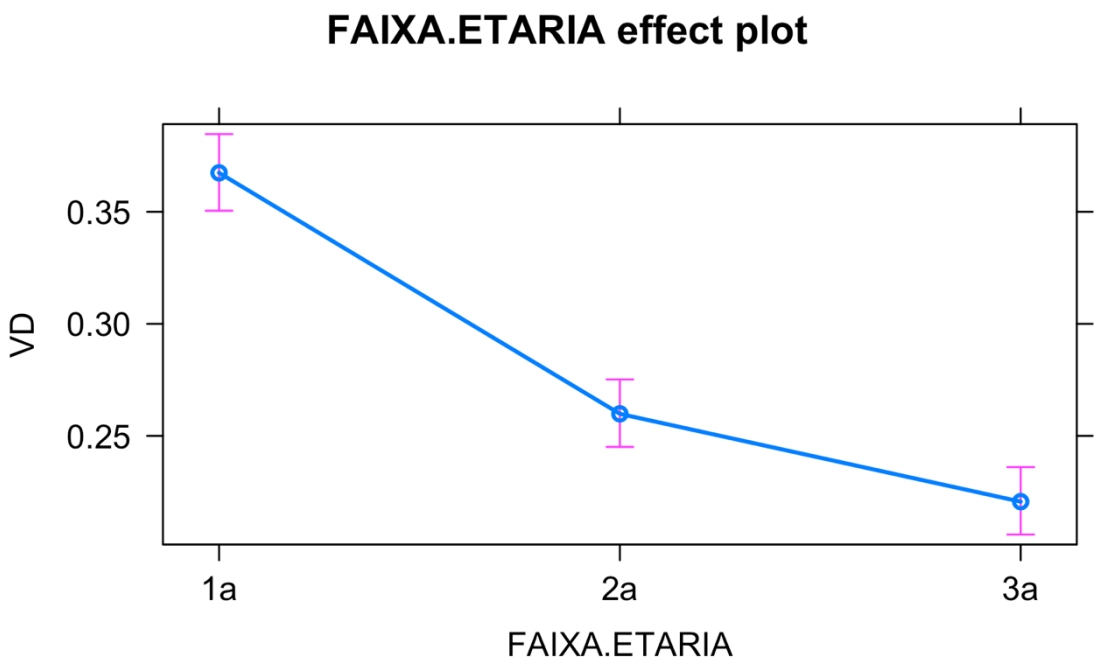

Figura 14.3: Gráfico de efeitos da variável Faixa Etária para o uso de /r/ retroflexo. Fonte: própria.

O gráfico de efeitos mostra claramente que, quanto mais velho o falante, menor a tendência a empregar retroflexo. Pela figura, também parece haver uma diferença significativa entre falantes de 2ª e 3ª faixas etárias.

Aplique a função `lrm()` ao mesmo modelo para obter outras medidas estatísticas relevantes.

```
lrm(VD ~ FAIXA.ETARIA, data = dados)

## Logistic Regression Model
##
##  lrm(formula = VD ~ FAIXA.ETARIA, data = dados)
##
##                       Model Likelihood     Discrimination    Rank Discri
m.
##                             Ratio Test            Indexes          Index
es
##Obs    9226      LR chi2     169.38    R2       0.026    C       0.5
80
##tepe  6615       d.f.             2    g        0.316    Dxy     0.1
59
##retr  2611       Pr(> chi2) <0.0001    gr       1.372    gamma   0.2
37
##max |deriv| 3e-11                      gp       0.065    tau-a   0.0
65
##                                       Brier    0.199
```

```
##
##                  Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept       -0.5433 0.0376 -14.47 <0.0001
##  FAIXA.ETARIA=2a -0.5036 0.0548  -9.18 <0.0001
##  FAIXA.ETARIA=3a -0.7187 0.0583 -12.32 <0.0001
##
```

Façamos agora um modelo com uma variável previsora numérica, INDICE.SOCIO. Guarde o resultado num objeto chamado `mod3`.

```
mod3 <- glm(VD ~ INDICE.SOCIO, data = dados, family = binomial)
```

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

```
summary(mod3)

##
## Call:
## glm(formula = VD ~ INDICE.SOCIO, family = binomial, data = dados)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2758  -0.8498  -0.6874   1.2902   2.0168
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.53407    0.12684   12.10   <2e-16 ***
## INDICE.SOCIO -0.81609    0.04201  -19.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10993  on 9225  degrees of freedom
## Residual deviance: 10596  on 9224  degrees of freedom
## AIC: 10600
##
## Number of Fisher Scoring iterations: 4
```

Como se trata de uma variável numérica contínua, a estimativa em logodds representa em quanto aumenta ou diminui a probabilidade de emprego de retroflexo a cada unidade de INDICE.SOCIO (ou seja, quanto mais alto o nível socioeconômico do falante). O coeficiente negativo -0,81609 indica que quanto mais alto o índice socioeconômico, menor a probabilidade de que o falante empregue o retroflexo.

Visualize esse resultado por meio de um gráfico de efeitos. A partir da última linha de comando em que se usou `plot()` e `allEffects()`, mude o nome do modelo para `mod3` (Figura 14.4).

```
plot(allEffects(mod3), type = “response”)
```

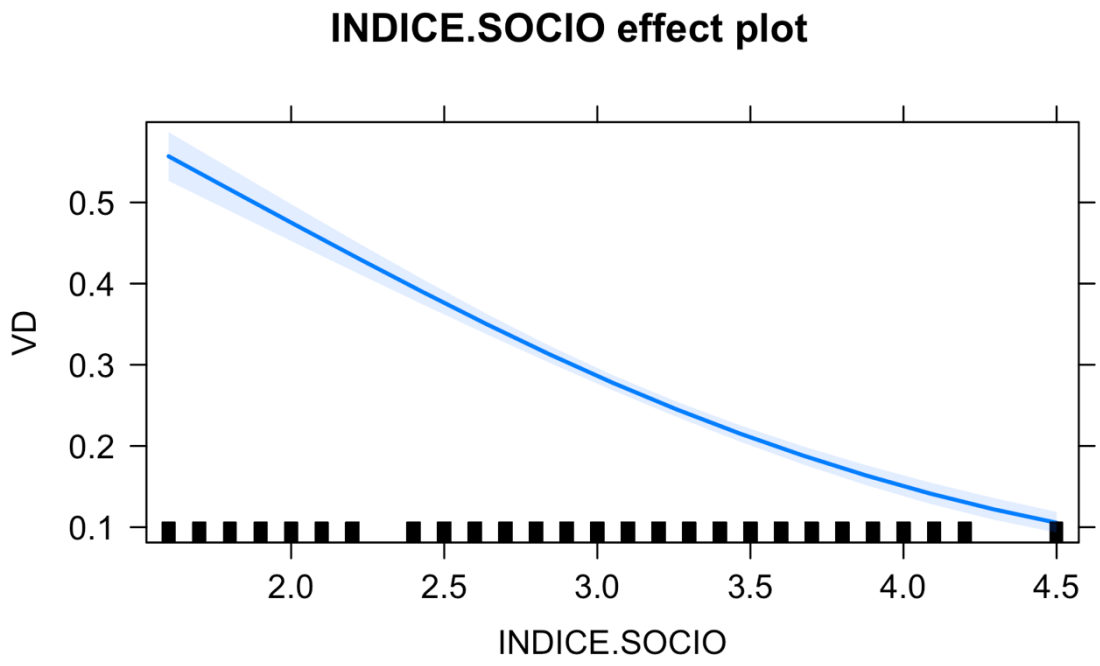

Figura 14.4: Gráfico de efeitos da variável Índice Socioeconômico para o uso de /r/ retroflexo. Fonte: própria.

Como já visto nos modelos lineares, a variável previsora numérica é representada pela linha de regressão. Vemos aí a drástica queda em probabilidade de emprego de retroflexo, de 50% para 10%, quanto mais alto o nível socioeconômico do falante.

Calcule a probabilidade, em logodds, de um falante com índice socioeconômico 4.2 empregar o retroflexo. Para isso, empregue os coeficientes gerados pelo modelo dentro da função y = 1.53407 + (-0.81609 * INDICE.SOCIO).

```
1.53407 + (-0.81609 * 4.2)
```

```
## [1] -1.893508
```

Transforme o valor em logodds em probabilidade. Para isso, aplique a função `ilogit()` à estimativa gerada acima.

```
ilogit(-1.893508)
```

```
## [1] 0.130845
```

Confira no gráfico de efeitos que a probabilidade calculada acima corresponde à medida do eixo y (VD) quando x = 4.2.

Agora obtenha as demais medidas estatísticas com uso da função `lrm()`.

```
lrm(VD ~ INDICE.SOCIO, data = dados)
```

```
## Logistic Regression Model
##
##  lrm(formula = VD ~ INDICE.SOCIO, data = dados)
##
##                    Model Likelihood    Discrimination    Rank Discri
m.
##                          Ratio Test           Indexes          Index
es
##Obs       9226    LR chi2     397.05    R2       0.061    C       0.
629
##tepe      6615    d.f.             1    g        0.535    Dxy     0.
259
##retr      2611    Pr(> chi2) <0.0001    gr       1.707    gamma   0.
273
##max |deriv| 3e-11                       gp       0.105    tau-a   0.
105
##                                        Brier    0.195
##
##               Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept     1.5341 0.1268  12.09 <0.0001
##  INDICE.SOCIO -0.8161 0.0420 -19.43 <0.0001
##
```

Vamos fazer agora um modelo com mais de uma variável previsora: VD ~ FAIXA.ETARIA + REGIAO. Chame esse modelo de `mod4`.

```
mod4 <- glm(VD ~ FAIXA.ETARIA + REGIAO, data = dados, family = binomia
l)
```

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

```
summary(mod4)
```

```
##
## Call:
## glm(formula = VD ~ FAIXA.ETARIA + REGIAO, family = binomial,
##     data = dados)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1156  -0.8393  -0.6064   1.2405   1.9638
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.05605    0.04841 -21.816   <2e-16 ***
## FAIXA.ETARIA2a   -0.54427    0.05606  -9.708   <2e-16 ***
## FAIXA.ETARIA3a   -0.71516    0.05949 -12.021   <2e-16 ***
## REGIAOperiferica  0.90882    0.04930  18.435   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10993  on 9225  degrees of freedom
## Residual deviance: 10467  on 9222  degrees of freedom
## AIC: 10475
##
## Number of Fisher Scoring iterations: 4
```

Vemos na tabela de coeficientes o mesmo resultado que havíamos visto para FAIXA.ETARIA (quanto mais velho, menor a tendência a usar retroflexo) e a correlação significativa com REGIAO (maior tendência ao retroflexo para os habitantes de bairros periféricos). Esse resultado não é surpreendente se levarmos em conta o resultado anterior para INDICE.SOCIO, uma vez que região de residência em São Paulo (e em muitas cidades) é um indicativo da classe social do falante.

Visualize os resultados com um gráfico de efeitos (Figura 14.5).

```
plot(allEffects(mod4), type = "response")
```

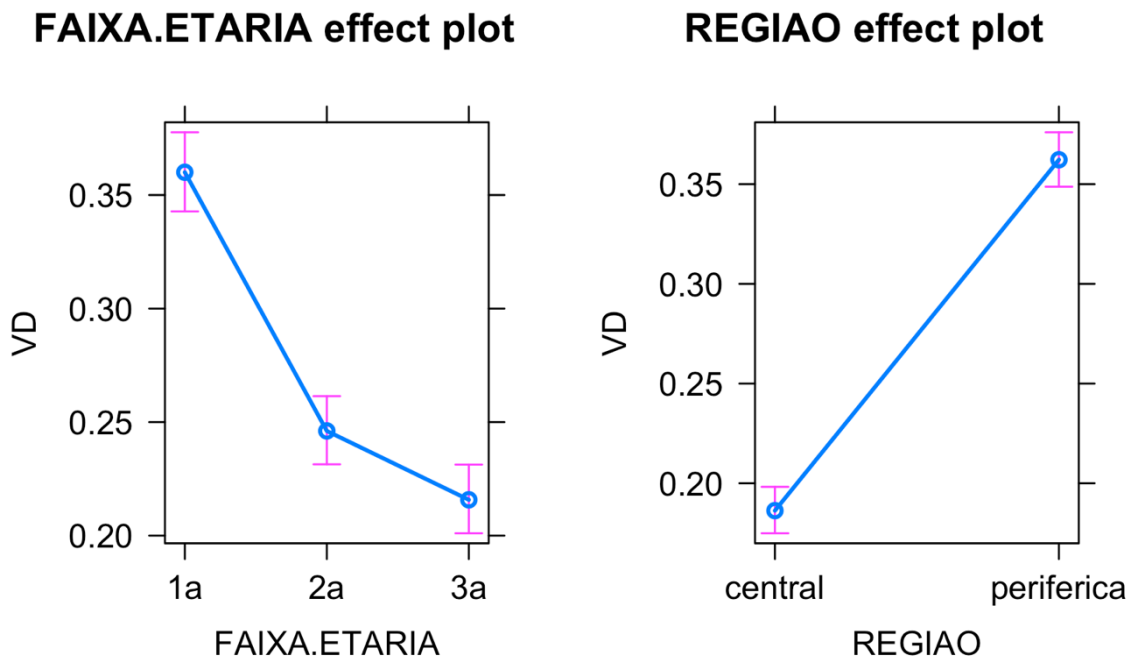

Figura 14.5: Gráfico de efeitos das variáveis Faixa Etária e Região de Residência para o uso de /r/ retroflexo. Fonte: própria.

E obtenha as demais estatísticas com a função `lrm()`.

```
lrm(VD ~ FAIXA.ETARIA + REGIAO, data = dados)
```

```
## Logistic Regression Model
##
##  lrm(formula = VD ~ FAIXA.ETARIA + REGIAO, data = dados)
##
##                    Model Likelihood    Discrimination    Rank Discri
m.
##                          Ratio Test           Indexes          Index
es
##Obs       9226    LR chi2     526.35    R2       0.080    C       0.
648
##tepe      6615    d.f.             3    g        0.616    Dxy     0.
296
##retr      2611    Pr(> chi2) <0.0001    gr       1.851    gamma   0.
352
##max |deriv| 9e-10                       gp       0.119    tau-a   0.
120
##                                        Brier    0.191
##
##                     Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept          -1.0560 0.0484 -21.82 <0.0001
##  FAIXA.ETARIA=2a    -0.5443 0.0561  -9.71 <0.0001
##  FAIXA.ETARIA=3a    -0.7152 0.0595 -12.02 <0.0001
##  REGIAO=periferica   0.9088 0.0493  18.44 <0.0001
##
```

Perdoe-me a repetição, mas é justamente isso que vai torná-lo proficiente em R! Façamos um último modelo simples, agora com interação.

A partir do último modelo criado com `glm()` com as variáveis FAIXA.ETARIA e REGIAO, mude o operador de soma para a multiplicação para testar se essas variáveis interagem entre si. Guarde o resultado num objeto chamado `mod5`.

```
mod5 <- glm(VD ~ FAIXA.ETARIA * REGIAO, data = dados, family = binomia
l)
```

Visualize o resultado de `mod5`.

```
summary(mod5)
```

```
##
## Call:
## glm(formula = VD ~ FAIXA.ETARIA * REGIAO, family = binomial,
##     data = dados)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1847  -0.8046  -0.6561   1.1701   1.8970
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|
)
## (Intercept)                     -1.31754    0.06490 -20.301  < 2e-1
```

```
6
## FAIXA.ETARIA2a                   -0.10887    0.09287  -1.172  0.2411
2
## FAIXA.ETARIA3a                   -0.30091    0.09562  -3.147  0.0016
5
## REGIAOperiferica                  1.33476    0.08168  16.341  < 2e-1
6
## FAIXA.ETARIA2a:REGIAOperiferica -0.69440    0.11688  -5.941 2.83e-0
9
## FAIXA.ETARIA3a:REGIAOperiferica -0.67794    0.12268  -5.526 3.27e-0
8
##
## (Intercept)                      ***
## FAIXA.ETARIA2a
## FAIXA.ETARIA3a                   **
## REGIAOperiferica                 ***
## FAIXA.ETARIA2a:REGIAOperiferica ***
## FAIXA.ETARIA3a:REGIAOperiferica ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10993  on 9225  degrees of freedom
## Residual deviance: 10421  on 9220  degrees of freedom
## AIC: 10433
##
## Number of Fisher Scoring iterations: 4
```

Nesse modelo com interação, a diferença entre a 1ª e a 2ª faixas etárias deixou de ser significativa (mas mantém-se a diferença entre a 1ª e a 3ª). A região periférica tem um coeficiente ainda mais alto do que em `mod4`, em favorecimento do retroflexo. Além disso, a interação entre periferia e as duas faixas etárias mais velhas é significativa; tais interações indicam que é necessário diminuir a estimativa para falantes de 2ª e 3ª faixas etárias da periferia para corrigi-lo em relação a um modelo sem interação.

Calcule a estimativa de emprego de retroflexo para um paulistano de 1a faixa etária que mora na periferia.

```
-1.31754 + 1.33476
```

```
## [1] 0.01722
```

Calcule a estimativa de emprego de retroflexo de um paulistano da 2a faixa etária que mora na região periférica. Note que, além das estimativas desses níveis, é necessário somar a estimativa do termo de interação.

```
-1.31754 -0.10887 + 1.33476 -0.69440
```

```
## [1] -0.78605
```

Com uma interação, fica ainda mais difícil interpretar os resultados numéricos, não? O melhor é *ver* o resultado. Faça um gráfico de efeitos de `mod5` para visualizar essas estimativas na escala de probabilidade (Figura 14.6).

```
plot(allEffects(mod5), type = "response")
```

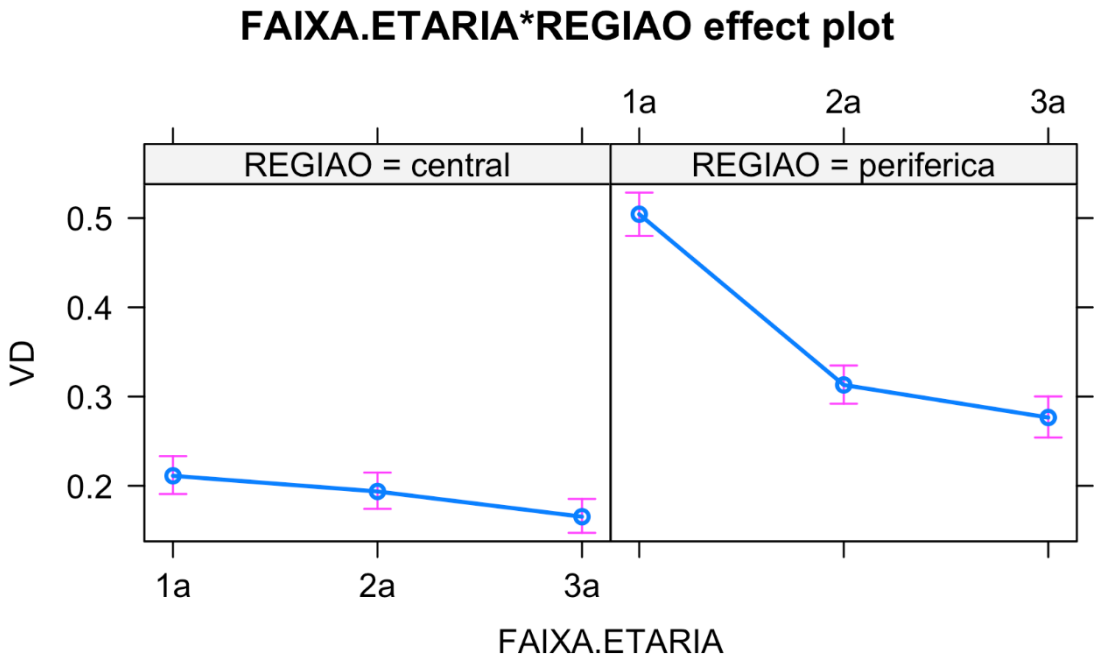

Figura 14.6: Gráfico de efeitos da interação entre as variáveis Faixa Etária e Região de Residência para o uso de /r/ retroflexo. Fonte: própria.

O gráfico de efeitos da interação mostra mais claramente que uma mudança na direção do retroflexo está ocorrendo na região periférica de São Paulo. Os falantes da região central, diferentemente, apresentam uma relativa estabilidade, com uma tendência bem mais tímida na direção do retroflexo. São os jovens de periferia que têm liderado essa possível mudança na comunidade. O fato de que os jovens da região central e da região periférica se comportam diferentemente é justamente o que caracteriza a interação: não é possível afirmar que há uma mudança na direção do retroflexo simplesmente: isso depende da região de residência do falante. Em outras palavras: FAIXA.ETARIA e REGIAO não são independentes, mas interagem entre si.

Aplique por fim a função `lrm()` ao modelo com interação para obter outras medidas estatísticas.

```
lrm(VD ~ FAIXA.ETARIA * REGIAO, data = dados)

## Logistic Regression Model
##
##  lrm(formula = VD ~ FAIXA.ETARIA * REGIAO, data = dados)
##
##                    Model Likelihood    Discrimination    Rank Discri
m.
##                          Ratio Test           Indexes          Index
es
##Obs       9226    LR chi2     571.97    R2       0.086    C       0.
648
##tepe      6615    d.f.             5    g        0.591    Dxy     0.
296
##retr      2611    Pr(> chi2) <0.0001    gr       1.806    gamma   0.
352
##max |deriv| 2e-09                       gp       0.120    tau-a   0.
120
##                                        Brier    0.190
##
##                                     Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept                          -1.3175 0.0649 -20.30 <0.0001
##  FAIXA.ETARIA=2a                    -0.1089 0.0929  -1.17 0.2411
##  FAIXA.ETARIA=3a                    -0.3009 0.0956  -3.15 0.0016
##  REGIAO=periferica                   1.3348 0.0817  16.34 <0.0001
##  FAIXA.ETARIA=2a * REGIAO=periferica -0.6944 0.1169  -5.94 <0.0001
##  FAIXA.ETARIA=3a * REGIAO=periferica -0.6779 0.1227  -5.53 <0.0001
##
```

Ao final do *script*, deixei os comandos para criar gráficos de efeitos de nossos modelos com funções do ggplot2, assim como as funções `logit()` e `ilogit()` que usamos para converter entre as medidas de probabilidade em diferentes escalas.

Na próxima lição, veremos modelos de regressão logística mais complexos e a aplicação de modelos de efeitos mistos a variáveis nominais binárias.

Para saber mais

Para saber mais sobre regressão logística e outros modelos aplicáveis a variáveis nominais, veja os capítulos 12 e 13 de Levshina (2015) e o capítulo 5 de Gries (2019).

Exercícios

Nesta lista de exercícios, você vai desenvolver uma análise semelhante à que fizemos na Lição 14, mas agora com outras variáveis previsoras. Primeiro, carregue os dados da planilha DadosRT.csv. Após carregar o dataframe, cheque sua estrutura, como de praxe. Certifique-se de que "retroflexo" é o segundo nível da VD, e organize variáveis ordinais em ordem lógica, de menor para maior. Para a variável REGIAO, defina a região central como o primeiro nível.

1. Faça um modelo de regressão logística para testar se há correlação entre a pronúncia de /r/ em coda (VD) e a posição da sílaba com /r/ na palavra (POSICAO.R). Há correlação significativa? Justifique sua resposta.
2. A qual nível de POSICAO.R se refere o coeficiente linear (Intercept)?
    a. posição medial
    b. posição final
3. A probabilidade de empregar retroflexo quando está no meio da palavra é...
    a. menor do que quando está no final da palavra
    b. maior do que quando está no final da palavra
4. Qual é o valor de índice C do modelo VD ~ POSICAO.R?
5. De acordo com a classificação de Hosmer & Lemeshow (2000, apud Levshina, 2015, p. 259), tal índice C tem...
    a. pouco poder de discriminação de resultado
    b. poder aceitável de discriminação de resultado
    c. poder excelente de discriminação de resultado
    d. poder notório de discriminação de resultado
6. Faça um modelo para testar se há correlação entre a pronúncia variável de /r/ em coda (VD) e a ESCOLARIDADE do falante. De acordo com o resultado deste modelo, não há diferença significativa entre...
    a. falantes com Ens. Fundamental e Ens.Médio
    b. falantes com Ens. Fundamental e Ens.Superior
    c. falantes com Ens. Médio e Ens. Superior

7. Em relação aos falantes menos escolarizados, os falantes com nível superior de escolaridade...
    a. tendem a empregar menos o retroflexo
    b. tendem a empregar mais o retroflexo
8. Transforme a medida de logodds da estimativa de emprego de retroflexo para falantes com nível superior para a medida de probabilidade (= proporção de 0% a 100%). Qual é a probabilidade de emprego de retroflexo para esses falantes?
    a. 21,9%
    b. 22,3%
    c. 33,9%
    d. 34,1%
    e. 50%
9. Faça um modelo para testar se há correlação entre a pronúncia variável de /r/ em coda (VD) e a IDADE do falante. A partir dele, calcule a estimativa, em logodds, de um falante com 50 anos de idade empregar o retroflexo.
10. Transforme o valor de logodds calculado na questão 9 para um valor de probabilidade.
11. Nesta lição, verificamos a interação entre FAIXA.ETARIA e REGIAO de residência do falante. A variável IDADE nada mais é do que a variável FAIXA.ETARIA vista de modo contínuo. Faça um modelo para testar a interação entre IDADE e REGIAO quanto ao uso variável de /r/ em coda. Há interação entre essas variáveis? Justifique sua resposta.
12. Calcule a probabilidade, em logodds, de um falante de 30 anos que mora na região central empregar o retroflexo.
13. Calcule a probabilidade, em logodds, de um falante de 30 anos que mora na região periférica empregar o retroflexo.
14. Transforme esta última medida de logodds em probabilidade.