Lição 15: Regressão Logística Parte 2

N.B.: Rode as linhas de comando a seguir antes de fazer 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))
  }
```

Nesta lição, vamos continuar trabalhando sobre o conjunto de dados da pronúncia variável de /r/ em coda silábica na fala de paulistanos. Como você já sabe, é uma boa ideia carregar os pacotes necessários para a sessão. Carregue os pacotes rms, effects, car, lme4 e lmerTest.

```
library(rms)
library(effects)
library(car)
library(lme4)
library(lmerTest)
```

No dataframe `dados`, o nível de referência de VD já foi redefinido, de modo que *tepe* é o primeiro nível e *retroflexo* é o segundo – e é aquele de acordo com o qual os resultados devem ser lidos (ver Lição 14).

Imagine que um pesquisador tenha levantado a hipótese de que a pronúncia de /r/ em coda (VD) se correlaciona com as variáveis SEXO.GENERO, FAIXA.ETARIA, REGIAO, INDICE.SOCIO, CONT.FON.PREC, CONT.FON.SEG, TONICIDADE, POSICAO.R e CLASSE.MORFOLOGICA. Aplique a função `str()` a `dados` para checar os níveis dessas variáveis previsoras.

```
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 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>
```

Algumas dessas variáveis não são totalmente ortogonais entre si e apresentam o risco de multicolinearidade (ver Lição 13). Duas variáveis são ortogonais quando todos os níveis de uma coocorrem com todos os níveis da outra. De modo simples, todas as combinações entre os níveis são possíveis. No entanto, este não é o caso para CONT.FON.SEG e POSICAO.R; essas variáveis não são totalmente ortogonais pois o contexto seguinte "pausa" só ocorre em final de palavra, nunca no meio. Faça uma tabela de frequências entre essas duas variáveis para visualizar isso. Dessa vez, vamos usar as funções do pacote base, `with()` e `table()`.

```
with(dados, table(CONT.FON.SEG, POSICAO.R))
```

```
##             POSICAO.R
## CONT.FON.SEG final medial
##      pausa     725      0
##      coronal   742   4090
##      dorsal    134   1908
##      labial    243   1384
```

As variáveis CLASSE.MORFOLOGICA e POSICAO.R também não são ortogonais entre si. Faça uma tabela de frequências entre elas para verificar a combinação que não ocorre.

```
with(dados, table(CLASSE.MORFOLOGICA, POSICAO.R))
```

```
##                   POSICAO.R
## CLASSE.MORFOLOGICA final medial
##        adjetivo      204   1326
##        adverbio       14     23
##        conj.prep      44    925
##        morf.inf      683      0
##        substantivo   766   4060
##        verbo         133   1048
```

Na codificação de CLASSE.MORFOLOGICA, foi feita uma separação entre o /r/ de infinitivo dos verbos (p.ex., “amar”) e o /r/ que ocorre na raiz da palavra (p.ex., “ergue”). O /r/ infinitivo no português só ocorre em final de palavra, nunca no meio. Ao mesmo tempo, o /r/ infinitivo é sempre tônico, de modo que CLASSE.MORFOLOGICA também não é totalmente ortogonal a TONICIDADE. Faça uma tabela de frequências entre CLASSE.MORFOLOGICA e TONICIDADE para visualizar isso.

```
with(dados, table(CLASSE.MORFOLOGICA, TONICIDADE))
```

```
##                   TONICIDADE
## CLASSE.MORFOLOGICA atona tonica
##        adjetivo      579    951
##        adverbio       21     16
##        conj.prep     868    101
##        morf.inf        0    683
##        substantivo  2302   2524
##        verbo         845    336
```

Tabelas de frequências como essas devem ser feitas sempre que se suspeita ou se prevê que duas variáveis não são ortogonais entre si. A presença de células vazias ou com poucos dados é o principal causador de multicolinearidade, o que, como visto na Lição 13, viola os pressupostos de modelos de regressão. Contudo, a depender do grau de colinearidade e do número de dados de que se dispõe, as (poucas) células vazias podem não ser um problema.

Faça um modelo de regressão logística `mod` que inclui as variáveis previsoras TONICIDADE, POSICAO.R, CLASSE.MORFOLOGICA e CONT.FON.SEG, em relação a VD.

```
mod <- glm(VD ~ TONICIDADE + POSICAO.R + CLASSE.MORFOLOGICA + CONT.FON
.SEG, data = dados, family = binomial)
```

Na Lição 13, vimos uma função que permite avaliar se há colinearidade entre variáveis. Qual é ela?

- `crPlot()`, do pacote car
- `effect()`, do pacote effects
- `lmer()`, do pacote lme4
- `lrm()`, do pacote rms
- `vif()`, do pacote car

Aplique então a função `vif()` a `mod`. Digite `car::vif(mod)`. (A notação pacote::funcao() é outro modo de disponibilizar uma biblioteca no R.)

```
car::vif(mod)
```

```
##                         GVIF Df GVIF^(1/(2*Df))
## TONICIDADE          1.356133  1        1.164531
## POSICAO.R           1.880253  1        1.371223
## CLASSE.MORFOLOGICA 1.658880  5        1.051917
## CONT.FON.SEG        1.640220  3        1.085968
```

O que informa o resultado de `vif()` acima?

- há colinearidade entre as variáveis
- não há colinearidade entre as variáveis

Apesar das células vazias, o conjunto de dados é robusto o suficiente para superar a potencial multicolinearidade entre essas variáveis. Podemos então seguir com a criação de um modelo mais complexo, com a inclusão de todas as variáveis mencionadas acima.

Na Lição 13, vimos que há duas abordagens principais para decidir a manutenção ou não de variáveis previsoras num modelo: uma "de baixo para cima", que começa com um modelo sem variáveis previsoras e tenta adicioná-las uma a uma, e outra "de cima para baixo", que começa com um modelo completo e tenta eliminar variáveis uma a uma. Ambas se implementam com a função `step()`; a primeira com direction = "forward", e a segunda com direction = "backward".

No *script* temos um primeiro modelo completo com todas as variáveis que queremos testar. Ele já está pronto, mas você deve se atentar à sua estrutura, para reproduzir a aplicação com seus dados.

```
modelo <- glm(VD ~
                SEXO.GENERO +
                FAIXA.ETARIA * REGIAO +
                INDICE.SOCIO +
                CONT.FON.PREC +
                CONT.FON.SEG +
                TONICIDADE +
                POSICAO.R +
                CLASSE.MORFOLOGICA,
              data = dados, family = binomial)
```

Veja o resultado com `summary()`.

```
summary(modelo)

##
## Call:
## glm(formula = VD ~ SEXO.GENERO + FAIXA.ETARIA * REGIAO + INDICE.SOC
IO +
##     CONT.FON.PREC + CONT.FON.SEG + TONICIDADE + POSICAO.R + CLASSE.
MORFOLOGICA,
##     family = binomial, data = dados)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1420  -0.7736  -0.4927   0.8614   2.9249
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|
)
## (Intercept)                     1.09412    0.23689   4.619 3.86e-0
6
## SEXO.GENEROmasculino            0.54806    0.05273  10.393  < 2e-1
6
## FAIXA.ETARIA2a                 -0.14534    0.09925  -1.464 0.14311
1
## FAIXA.ETARIA3a                 -0.45935    0.10264  -4.475 7.62e-0
6
## REGIAOperiferica                1.28609    0.08849  14.534  < 2e-1
6
## INDICE.SOCIO                   -0.90138    0.04828 -18.671  < 2e-1
6
## CONT.FON.PRECe                  0.31751    0.13068   2.430 0.01510
8
## CONT.FON.PREC3                  0.52761    0.15575   3.388 0.00070
5
## CONT.FON.PRECa                  0.50403    0.13190   3.821 0.00013
3
## CONT.FON.PREC0                  0.47885    0.15456   3.098 0.00194
8
## CONT.FON.PRECo                 -0.31598    0.13825  -2.286 0.02227
7
## CONT.FON.PRECu                 -0.95710    0.17397  -5.502 3.76e-0
8
## CONT.FON.SEGcoronal             0.67861    0.11960   5.674 1.40e-0
8
## CONT.FON.SEGdorsal             -0.21684    0.13535  -1.602 0.10914
4
## CONT.FON.SEGlabial              0.23678    0.13011   1.820 0.06878
0
## TONICIDADEtonica                0.29545    0.07009   4.215 2.49e-0
5
## POSICAO.Rmedial                -0.82823    0.08958  -9.246  < 2e-1
6
## CLASSE.MORFOLOGICAadverbio      1.04562    0.37562   2.784 0.00537
4
## CLASSE.MORFOLOGICAconj.prep     0.47925    0.15094   3.175 0.00149
```

```
8
## CLASSE.MORFOLOGICAmorf.inf      -1.10812    0.14356  -7.719 1.17e-1
4
## CLASSE.MORFOLOGICAsubstantivo    0.12488    0.07681   1.626 0.10397
0
## CLASSE.MORFOLOGICAverbo          0.47789    0.09904   4.825 1.40e-0
6
## FAIXA.ETARIA2a:REGIAOperiferica -0.71666    0.12544  -5.713 1.11e-0
8
## FAIXA.ETARIA3a:REGIAOperiferica -0.76762    0.13267  -5.786 7.21e-0
9
##
## (Intercept)                     ***
## SEXO.GENEROmasculino            ***
## FAIXA.ETARIA2a
## FAIXA.ETARIA3a                  ***
## REGIAOperiferica                ***
## INDICE.SOCIO                    ***
## CONT.FON.PRECe                  *
## CONT.FON.PREC3                  ***
## CONT.FON.PRECa                  ***
## CONT.FON.PREC0                  **
## CONT.FON.PRECo                  *
## CONT.FON.PRECu                  ***
## CONT.FON.SEGcoronal             ***
## CONT.FON.SEGdorsal
## CONT.FON.SEGlabial              .
## TONICIDADEtonica                ***
## POSICAO.Rmedial                 ***
## CLASSE.MORFOLOGICAadverbio      **
## CLASSE.MORFOLOGICAconj.prep     **
## CLASSE.MORFOLOGICAmorf.inf      ***
## CLASSE.MORFOLOGICAsubstantivo
## CLASSE.MORFOLOGICAverbo         ***
## 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: 10993.1  on 9225  degrees of freedom
## Residual deviance:  9137.3  on 9202  degrees of freedom
## AIC: 9185.3
##
## Number of Fisher Scoring iterations: 5
```

Para interpretar os resultados, é necessário saber o nível de referência de cada variável. Qual é o nível de referência da variável SEXO?

- feminino
- masculino

Qual é o nível de referência da variável FAIXA.ETARIA?

- 1a
- 2a
- 3a

Qual é o nível de referência da variável REGIAO?

- central
- periferica

Qual é o nível de referência da variável INDICE.SOCIO?

- INDICE.SOCIO = 0
- INDICE.SOCIO = 5

Qual é o nível de referência da variável CONT.FON.PREC?

- a
- e
- i
- o
- u
- 3
- 0

Para a variável CONT.FON.PREC, os níveis já foram reorganizados no conjunto `dados`. Em vez da ordem alfabética, as vogais que precedem /r/ em coda foram ordenadas de [+anterior] para [-anterior], e [+alta] para [-alta]. Lembre-se que os níveis '3' e '0' correspondem às vogais média baixas /ɛ/ e /ɔ/ respectivamente.

Qual é o nível de referência da variável CONT.FON.SEG?

- coronal
- dorsal
- labial
- pausa

Assim como a variável CONT.FON.PREC, CONT.FON.SEG foi reorganizada previamente, no momento de importação dos dados. A variável foi originalmente codificada de modo detalhado, de acordo com cada segmento consonantal (/b/, /d/, /g/ etc.) ou existência de pausa, mas os segmentos foram recodificados aqui pelo Ponto de C (ver Clements & Hume, 1995). A pausa foi mantida e definida como nível de referência.

Qual é o nível de referência da variável TONICIDADE?

- atona
- tonica

Qual é o nível de referência da variável POSICAO.R?

- final
- medial

Qual é o nível de referência da variável CLASSE.MORFOLOGICA?

- adjetivo
- adverbio
- conj.prep
- morf.inf
- substantivo
- verbo

Trata-se de um modelo bastante complexo! O resultado para o coeficiente linear (Intercept), então, refere-se à pronúncia como *retroflexo* – lembre-se que os resultados são lidos em relação ao segundo nível da VD, que foi redefinida –, na fala de mulheres de 1ª faixa etária residentes da região central e de nível socioeconômico zero, em ocorrências de /r/ em coda precedidas da vogal /i/, seguidas de pausa, em sílabas átonas, em posição final de adjetivos! (respira... respira...) Veja que a leitura e interpretação de resultados fica mais difícil à medida que se incluem novas variáveis previsoras, de modo que deve haver bons motivos para incluí-las no modelo.

Aplique a função `lrm()` para obter as demais medidas estatísticas do modelo. Lembre-se que esta função toma como argumentos a fórmula e o conjunto de dados (sem family = binomial).

```
lrm(VD ~
      SEXO.GENERO +
      FAIXA.ETARIA * REGIAO +
      INDICE.SOCIO +
      CONT.FON.PREC +
      CONT.FON.SEG +
      TONICIDADE +
      POSICAO.R +
      CLASSE.MORFOLOGICA,
    data = dados)
```

```
## Logistic Regression Model
##
##  lrm(formula = VD ~ SEXO.GENERO + FAIXA.ETARIA * REGIAO + INDICE.SO
CIO
##      + CONT.FON.PREC + CONT.FON.SEG + TONICIDADE + POSICAO.R + CLAS
SE.  MORFOLOGICA,
##      data = dados)
##
##                    Model Likelihood    Discrimination    Rank Discri
m.
##                          Ratio Test           Indexes          Index
es
##Obs      9226    LR chi2    1855.86    R2       0.262    C       0.7
74
##tepe     6615    d.f.            23    g        1.333    Dxy     0.5
49
##retrofl  2611    Pr(> chi2) <0.0001    gr       3.793    gamma   0.5
49
##max |deriv| 2e-08                      gp       0.221    tau-a   0.2
23
##                                       Brier    0.163
##
##                                      Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept                            1.0941 0.2369   4.62 <0.0001
##  SEXO.GENERO=masculino                0.5481 0.0527  10.39 <0.0001
##  FAIXA.ETARIA=2a                     -0.1453 0.0993  -1.46 0.1431
##  FAIXA.ETARIA=3a                     -0.4594 0.1026  -4.48 <0.0001
##  REGIAO=periferica                    1.2861 0.0885  14.53 <0.0001
##  INDICE.SOCIO                        -0.9014 0.0483 -18.67 <0.0001
##  CONT.FON.PREC=e                      0.3175 0.1307   2.43 0.0151
##  CONT.FON.PREC=3                      0.5276 0.1557   3.39 0.0007
##  CONT.FON.PREC=a                      0.5040 0.1319   3.82 0.0001
##  CONT.FON.PREC=0                      0.4788 0.1546   3.10 0.0019
##  CONT.FON.PREC=o                     -0.3160 0.1382  -2.29 0.0223
##  CONT.FON.PREC=u                     -0.9571 0.1740  -5.50 <0.0001
##  CONT.FON.SEG=coronal                 0.6786 0.1196   5.67 <0.0001
##  CONT.FON.SEG=dorsal                 -0.2168 0.1354  -1.60 0.1091
```

```
##   CONT.FON.SEG=labial                     0.2368 0.1301   1.82 0.0688
##   TONICIDADE=tonica                       0.2954 0.0701   4.22 <0.0001
##   POSICAO.R=medial                       -0.8282 0.0896  -9.25 <0.0001
##   CLASSE.MORFOLOGICA=adverbio             1.0456 0.3756   2.78 0.0054
##   CLASSE.MORFOLOGICA=conj.prep            0.4792 0.1509   3.18 0.0015
##   CLASSE.MORFOLOGICA=morf.inf            -1.1081 0.1436  -7.72 <0.0001
##   CLASSE.MORFOLOGICA=substantivo          0.1249 0.0768   1.63 0.1040
##   CLASSE.MORFOLOGICA=verbo                0.4779 0.0990   4.83 <0.0001
##   FAIXA.ETARIA=2a * REGIAO=periferica -0.7167 0.1254  -5.71 <0.0001
##   FAIXA.ETARIA=3a * REGIAO=periferica -0.7676 0.1327  -5.79 <0.0001
##
```

Lembra que os modelos da lição anterior, com apenas uma ou duas variáveis previsoras, tinham índice C baixos? Nosso modelo agora está num nível aceitável, com C = 0,774. Mas será que todas as variáveis são relevantes?

Crie agora um modelo sem qualquer variável previsora, a partir do qual faremos o step forward. Digite `m0 <- glm(VD ~ 1, data = dados, family = binomial)`.

```
m0 <- glm(VD ~ 1, data = dados, family = binomial)
```

Na Lição 13, vimos que a função `step()`, na direção "forward", toma como argumentos: (i) um modelo sem variáveis previsoras – aqui, `m0`; (ii) direction = "forward"; e scope com ~ (sem a Variável Resposta) e a especificação de variáveis.

```
m.fw <- step(m0,
             direction = "forward",
             scope = ~
               SEXO.GENERO +
               FAIXA.ETARIA * REGIAO +
               INDICE.SOCIO +
               CONT.FON.PREC +
               CONT.FON.SEG +
               TONICIDADE +
               POSICAO.R +
               CLASSE.MORFOLOGICA)

## Start:  AIC=10995.12
## VD ~ 1
##
##                       Df Deviance   AIC
## + CONT.FON.PREC        6    10532 10546
## + INDICE.SOCIO         1    10596 10600
## + REGIAO               1    10635 10639
## + CONT.FON.SEG         3    10724 10732
## + CLASSE.MORFOLOGICA   5    10792 10804
## + FAIXA.ETARIA         2    10824 10830
## + TONICIDADE           1    10847 10851
## + SEXO.GENERO          1    10903 10907
## + POSICAO.R            1    10920 10924
```

```
## <none>                     10993 10995
##
## Step:  AIC=10546.06
## VD ~ CONT.FON.PREC
##
##                       Df Deviance   AIC
## + INDICE.SOCIO         1    10126 10142
## + REGIAO               1    10161 10177
## + FAIXA.ETARIA         2    10364 10382
## + CONT.FON.SEG         3    10390 10410
## + SEXO.GENERO          1    10451 10467
## + CLASSE.MORFOLOGICA   5    10476 10500
## + POSICAO.R            1    10497 10513
## + TONICIDADE           1    10510 10526
## <none>                      10532 10546
##
## Step:  AIC=10141.82
## VD ~ CONT.FON.PREC + INDICE.SOCIO
##
##                       Df Deviance     AIC
## + REGIAO               1   9864.8  9882.8
## + FAIXA.ETARIA         2   9891.3  9911.3
## + CONT.FON.SEG         3   9988.5 10010.5
## + SEXO.GENERO          1  10008.9 10026.9
## + CLASSE.MORFOLOGICA   5  10074.1 10100.1
## + POSICAO.R            1  10091.1 10109.1
## + TONICIDADE           1  10098.4 10116.4
## <none>                    10125.8 10141.8
##
## Step:  AIC=9882.75
## VD ~ CONT.FON.PREC + INDICE.SOCIO + REGIAO
##
##                       Df Deviance    AIC
## + FAIXA.ETARIA         2   9646.5 9668.5
## + CONT.FON.SEG         3   9720.5 9744.5
## + SEXO.GENERO          1   9752.4 9772.4
## + CLASSE.MORFOLOGICA   5   9814.5 9842.5
## + POSICAO.R            1   9827.8 9847.8
## + TONICIDADE           1   9840.2 9860.2
## <none>                     9864.8 9882.8
##
## Step:  AIC=9668.49
## VD ~ CONT.FON.PREC + INDICE.SOCIO + REGIAO + FAIXA.ETARIA
##
##                        Df Deviance    AIC
## + CONT.FON.SEG          3   9504.4 9532.4
## + SEXO.GENERO           1   9539.4 9563.4
## + CLASSE.MORFOLOGICA    5   9596.4 9628.4
## + FAIXA.ETARIA:REGIAO   2   9603.1 9629.1
## + POSICAO.R             1   9606.8 9630.8
## + TONICIDADE            1   9621.0 9645.0
## <none>                      9646.5 9668.5
##
## Step:  AIC=9532.4
```

```
## VD ~ CONT.FON.PREC + INDICE.SOCIO + REGIAO + FAIXA.ETARIA + CONT.FO
N.SEG
##
##                        Df Deviance    AIC
## + SEXO.GENERO           1   9401.0 9431.0
## + POSICAO.R             1   9439.3 9469.3
## + CLASSE.MORFOLOGICA    5   9438.5 9476.5
## + FAIXA.ETARIA:REGIAO   2   9461.3 9493.3
## + TONICIDADE            1   9476.4 9506.4
## <none>                      9504.4 9532.4
##
## Step:  AIC=9430.96
## VD ~ CONT.FON.PREC + INDICE.SOCIO + REGIAO + FAIXA.ETARIA + CONT.FO
N.SEG +
##     SEXO.GENERO
##
##                        Df Deviance    AIC
## + POSICAO.R             1   9332.5 9364.5
## + CLASSE.MORFOLOGICA    5   9336.2 9376.2
## + FAIXA.ETARIA:REGIAO   2   9354.8 9388.8
## + TONICIDADE            1   9371.8 9403.8
## <none>                      9401.0 9431.0
##
## Step:  AIC=9364.48
## VD ~ CONT.FON.PREC + INDICE.SOCIO + REGIAO + FAIXA.ETARIA + CONT.FO
N.SEG +
##     SEXO.GENERO + POSICAO.R
##
##                        Df Deviance    AIC
## + CLASSE.MORFOLOGICA    5   9201.0 9243.0
## + FAIXA.ETARIA:REGIAO   2   9286.6 9322.6
## + TONICIDADE            1   9328.6 9362.6
## <none>                      9332.5 9364.5
##
## Step:  AIC=9242.96
## VD ~ CONT.FON.PREC + INDICE.SOCIO + REGIAO + FAIXA.ETARIA + CONT.FO
N.SEG +
##     SEXO.GENERO + POSICAO.R + CLASSE.MORFOLOGICA
##
##                        Df Deviance    AIC
## + FAIXA.ETARIA:REGIAO   2   9154.9 9200.9
## + TONICIDADE            1   9183.0 9227.0
## <none>                      9201.0 9243.0
##
## Step:  AIC=9200.92
## VD ~ CONT.FON.PREC + INDICE.SOCIO + REGIAO + FAIXA.ETARIA + CONT.FO
N.SEG +
##     SEXO.GENERO + POSICAO.R + CLASSE.MORFOLOGICA + REGIAO:FAIXA.ETA
RIA
##
##              Df Deviance    AIC
## + TONICIDADE  1   9137.3 9185.3
## <none>            9154.9 9200.9
##
```

```
## Step:  AIC=9185.26
## VD ~ CONT.FON.PREC + INDICE.SOCIO + REGIAO + FAIXA.ETARIA + CONT.FO
N.SEG +
##     SEXO.GENERO + POSICAO.R + CLASSE.MORFOLOGICA + TONICIDADE +
##     REGIAO:FAIXA.ETARIA
```

O resultado é bem mais longo do que havíamos visto para o modelo de regressão linear porque aqui incluímos mais variáveis previsoras. O resultado de step forward indica que todas as variáveis devem ser mantidas no modelo.

Visualize também o resultado guardado em `m.fw`.

```
m.fw

##
## Call:  glm(formula = VD ~ CONT.FON.PREC + INDICE.SOCIO + REGIAO + F
AIXA.ETARIA +
##     CONT.FON.SEG + SEXO.GENERO + POSICAO.R + CLASSE.MORFOLOGICA +
##     TONICIDADE + REGIAO:FAIXA.ETARIA, family = binomial, data = dad
os)
##
## Coefficients:
##                       (Intercept)                    CONT.FON.PRECe
##                            1.0941                            0.3175
##                   CONT.FON.PREC3                    CONT.FON.PRECa
##                            0.5276                            0.5040
##                   CONT.FON.PREC0                    CONT.FON.PRECo
##                            0.4788                           -0.3160
##                   CONT.FON.PRECu                      INDICE.SOCIO
##                           -0.9571                           -0.9014
##                 REGIAOperiferica                    FAIXA.ETARIA2a
##                            1.2861                           -0.1453
##                   FAIXA.ETARIA3a               CONT.FON.SEGcoronal
##                           -0.4594                            0.6786
##               CONT.FON.SEGdorsal                CONT.FON.SEGlabial
##                           -0.2168                            0.2368
##             SEXO.GENEROmasculino                   POSICAO.Rmedial
##                            0.5481                           -0.8282
##      CLASSE.MORFOLOGICAadverbio      CLASSE.MORFOLOGICAconj.prep
##                            1.0456                            0.4792
##      CLASSE.MORFOLOGICAmorf.inf    CLASSE.MORFOLOGICAsubstantivo
##                           -1.1081                            0.1249
##         CLASSE.MORFOLOGICAverbo                  TONICIDADEtonica
##                            0.4779                            0.2954
## REGIAOperiferica:FAIXA.ETARIA2a  REGIAOperiferica:FAIXA.ETARIA3a
##                           -0.7167                           -0.7676
##
## Degrees of Freedom: 9225 Total (i.e. Null);  9202 Residual
## Null Deviance:       10990
## Residual Deviance: 9137  AIC: 9185
```

`m.fw` mostra os coeficientes angulares do modelo final. Façamos agora o modelo "de trás para frente", a fim de verificar se as mesmas variáveis previsoras são selecionadas. Digite `m.bw <- step(modelo, direction = "backward")`.

```
m.bw <- step(modelo, direction = "backward")

## Start:  AIC=9185.26
## VD ~ SEXO.GENERO + FAIXA.ETARIA * REGIAO + INDICE.SOCIO + CONT.FON.
PREC +
##     CONT.FON.SEG + TONICIDADE + POSICAO.R + CLASSE.MORFOLOGICA
##
##                        Df Deviance    AIC
## <none>                      9137.3 9185.3
## - TONICIDADE            1   9154.9 9200.9
## - FAIXA.ETARIA:REGIAO   2   9183.0 9227.0
## - POSICAO.R             1   9222.9 9268.9
## - SEXO.GENERO           1   9247.3 9293.3
## - CLASSE.MORFOLOGICA    5   9282.9 9320.9
## - CONT.FON.SEG          3   9304.3 9346.3
## - CONT.FON.PREC         6   9369.8 9405.8
## - INDICE.SOCIO          1   9508.3 9554.3
```

Assim como fizemos na Lição 13, aqui partimos do modelo completo (`modelo`) e tentamos eliminar variáveis. O resultado final recomenda a manutenção de todas as variáveis previsoras.

Visualize os coeficientes guardados em `m.bw`.

```
m.bw

##
## Call:  glm(formula = VD ~ SEXO.GENERO + FAIXA.ETARIA * REGIAO + IND
ICE.SOCIO +
##     CONT.FON.PREC + CONT.FON.SEG + TONICIDADE + POSICAO.R + CLASSE.
MORFOLOGICA,
##     family = binomial, data = dados)
##
## Coefficients:
##                     (Intercept)             SEXO.GENEROmasculino
##                          1.0941                          0.5481
##                 FAIXA.ETARIA2a                  FAIXA.ETARIA3a
##                         -0.1453                         -0.4594
##               REGIAOperiferica                    INDICE.SOCIO
##                          1.2861                         -0.9014
##                 CONT.FON.PRECe                  CONT.FON.PREC3
##                          0.3175                          0.5276
##                 CONT.FON.PRECa                  CONT.FON.PREC0
##                          0.5040                          0.4788
##                 CONT.FON.PRECo                  CONT.FON.PRECu
##                         -0.3160                         -0.9571
##            CONT.FON.SEGcoronal               CONT.FON.SEGdorsal
```

```
##                            0.6786                            -0.2168
##               CONT.FON.SEGlabial                 TONICIDADEtonica
##                            0.2368                             0.2954
##                   POSICAO.Rmedial       CLASSE.MORFOLOGICAadverbio
##                           -0.8282                             1.0456
##     CLASSE.MORFOLOGICAconj.prep       CLASSE.MORFOLOGICAmorf.inf
##                            0.4792                            -1.1081
##   CLASSE.MORFOLOGICAsubstantivo          CLASSE.MORFOLOGICAverbo
##                            0.1249                             0.4779
## FAIXA.ETARIA2a:REGIAOperiferica  FAIXA.ETARIA3a:REGIAOperiferica
##                           -0.7167                            -0.7676
##
## Degrees of Freedom: 9225 Total (i.e. Null);  9202 Residual
## Null Deviance:       10990
## Residual Deviance: 9137  AIC: 9185
```

O objetivo aqui é verificar se os coeficientes do modelo "para frente" coincidem com aqueles do modelo "de trás para frente". Com um pouco de paciência, verificamos que os coeficientes são os mesmos.

Aplique também a função `step()` com direction = "both". Recorde-se que neste caso a função `step()` se inicia como a direção "forward" mas, a cada variável incluída no modelo, ele tenta excluir alguma que não seja mais relevante. Os argumentos são os mesmos que na direção "forward", com a exclusão de direction (já que a opção "both" é a *default*).

```
m.both <- step(m0,
               scope = ~
                 SEXO.GENERO +
                 FAIXA.ETARIA * REGIAO +
                 INDICE.SOCIO +
                 CONT.FON.PREC +
                 CONT.FON.SEG +
                 TONICIDADE +
                 POSICAO.R +
                 CLASSE.MORFOLOGICA)

## Start:  AIC=10995.12
## VD ~ 1
##
##                       Df Deviance   AIC
## + CONT.FON.PREC        6    10532 10546
## + INDICE.SOCIO         1    10596 10600
## + REGIAO               1    10635 10639
## + CONT.FON.SEG         3    10724 10732
## + CLASSE.MORFOLOGICA   5    10792 10804
## + FAIXA.ETARIA         2    10824 10830
## + TONICIDADE           1    10847 10851
## + SEXO.GENERO          1    10903 10907
```

```
## + POSICAO.R           1    10920 10924
## <none>                     10993 10995
##
## Step:  AIC=10546.06
## VD ~ CONT.FON.PREC
##
##                      Df Deviance   AIC
## + INDICE.SOCIO        1    10126 10142
## + REGIAO              1    10161 10177
## + FAIXA.ETARIA        2    10364 10382
## + CONT.FON.SEG        3    10390 10410
## + SEXO.GENERO         1    10451 10467
## + CLASSE.MORFOLOGICA  5    10476 10500
## + POSICAO.R           1    10497 10513
## + TONICIDADE          1    10510 10526
## <none>                     10532 10546
## - CONT.FON.PREC       6    10993 10995
##
## Step:  AIC=10141.82
## VD ~ CONT.FON.PREC + INDICE.SOCIO
##
##                      Df Deviance     AIC
## + REGIAO              1   9864.8  9882.8
## + FAIXA.ETARIA        2   9891.3  9911.3
## + CONT.FON.SEG        3   9988.5 10010.5
## + SEXO.GENERO         1  10008.9 10026.9
## + CLASSE.MORFOLOGICA  5  10074.1 10100.1
## + POSICAO.R           1  10091.1 10109.1
## + TONICIDADE          1  10098.4 10116.4
## <none>                   10125.8 10141.8
## - INDICE.SOCIO        1  10532.1 10546.1
## - CONT.FON.PREC       6  10596.1 10600.1
##
## Step:  AIC=9882.75
## VD ~ CONT.FON.PREC + INDICE.SOCIO + REGIAO
##
##                      Df Deviance     AIC
## + FAIXA.ETARIA        2   9646.5  9668.5
## + CONT.FON.SEG        3   9720.5  9744.5
## + SEXO.GENERO         1   9752.4  9772.4
## + CLASSE.MORFOLOGICA  5   9814.5  9842.5
## + POSICAO.R           1   9827.8  9847.8
## + TONICIDADE          1   9840.2  9860.2
## <none>                    9864.8  9882.8
## - REGIAO              1  10125.8 10141.8
## - INDICE.SOCIO        1  10160.7 10176.7
## - CONT.FON.PREC       6  10347.8 10353.8
##
## Step:  AIC=9668.49
## VD ~ CONT.FON.PREC + INDICE.SOCIO + REGIAO + FAIXA.ETARIA
##
##                       Df Deviance     AIC
## + CONT.FON.SEG         3   9504.4  9532.4
## + SEXO.GENERO          1   9539.4  9563.4
```

```
## + CLASSE.MORFOLOGICA   5   9596.4  9628.4
## + FAIXA.ETARIA:REGIAO  2   9603.1  9629.1
## + POSICAO.R            1   9606.8  9630.8
## + TONICIDADE           1   9621.0  9645.0
## <none>                     9646.5  9668.5
## - FAIXA.ETARIA         2   9864.8  9882.8
## - REGIAO               1   9891.3  9911.3
## - INDICE.SOCIO         1   9993.5 10013.5
## - CONT.FON.PREC        6  10133.3 10143.3
##
## Step:  AIC=9532.4
## VD ~ CONT.FON.PREC + INDICE.SOCIO + REGIAO + FAIXA.ETARIA + CONT.FO
N.SEG
##
##                        Df Deviance    AIC
## + SEXO.GENERO           1   9401.0 9431.0
## + POSICAO.R             1   9439.3 9469.3
## + CLASSE.MORFOLOGICA    5   9438.5 9476.5
## + FAIXA.ETARIA:REGIAO   2   9461.3 9493.3
## + TONICIDADE            1   9476.4 9506.4
## <none>                      9504.4 9532.4
## - CONT.FON.SEG          3   9646.5 9668.5
## - FAIXA.ETARIA          2   9720.5 9744.5
## - REGIAO                1   9755.3 9781.3
## - CONT.FON.PREC         6   9853.6 9869.6
## - INDICE.SOCIO          1   9845.5 9871.5
##
## Step:  AIC=9430.96
## VD ~ CONT.FON.PREC + INDICE.SOCIO + REGIAO + FAIXA.ETARIA + CONT.FO
N.SEG +
##     SEXO.GENERO
##
##                        Df Deviance    AIC
## + POSICAO.R             1   9332.5 9364.5
## + CLASSE.MORFOLOGICA    5   9336.2 9376.2
## + FAIXA.ETARIA:REGIAO   2   9354.8 9388.8
## + TONICIDADE            1   9371.8 9403.8
## <none>                      9401.0 9431.0
## - SEXO.GENERO           1   9504.4 9532.4
## - CONT.FON.SEG          3   9539.4 9563.4
## - FAIXA.ETARIA          2   9611.7 9637.7
## - REGIAO                1   9646.2 9674.2
## - CONT.FON.PREC         6   9745.6 9763.6
## - INDICE.SOCIO          1   9781.4 9809.4
##
## Step:  AIC=9364.48
## VD ~ CONT.FON.PREC + INDICE.SOCIO + REGIAO + FAIXA.ETARIA + CONT.FO
N.SEG +
##     SEXO.GENERO + POSICAO.R
##
##                        Df Deviance    AIC
## + CLASSE.MORFOLOGICA    5   9201.0 9243.0
## + FAIXA.ETARIA:REGIAO   2   9286.6 9322.6
## + TONICIDADE            1   9328.6 9362.6
```

```
## <none>                     9332.5 9364.5
## - POSICAO.R            1   9401.0 9431.0
## - SEXO.GENERO          1   9439.3 9469.3
## - CONT.FON.SEG         3   9497.5 9523.5
## - FAIXA.ETARIA         2   9546.1 9574.1
## - REGIAO               1   9582.9 9612.9
## - CONT.FON.PREC        6   9648.0 9668.0
## - INDICE.SOCIO         1   9715.5 9745.5
##
## Step:  AIC=9242.96
## VD ~ CONT.FON.PREC + INDICE.SOCIO + REGIAO + FAIXA.ETARIA + CONT.FO
N.SEG +
##     SEXO.GENERO + POSICAO.R + CLASSE.MORFOLOGICA
##
##                         Df Deviance    AIC
## + FAIXA.ETARIA:REGIAO  2   9154.9 9200.9
## + TONICIDADE           1   9183.0 9227.0
## <none>                     9201.0 9243.0
## - SEXO.GENERO          1   9306.9 9346.9
## - CLASSE.MORFOLOGICA   5   9332.5 9364.5
## - POSICAO.R            1   9336.2 9376.2
## - CONT.FON.SEG         3   9368.5 9404.5
## - FAIXA.ETARIA         2   9416.9 9454.9
## - REGIAO               1   9451.2 9491.2
## - CONT.FON.PREC        6   9481.9 9511.9
## - INDICE.SOCIO         1   9573.0 9613.0
##
## Step:  AIC=9200.92
## VD ~ CONT.FON.PREC + INDICE.SOCIO + REGIAO + FAIXA.ETARIA + CONT.FO
N.SEG +
##     SEXO.GENERO + POSICAO.R + CLASSE.MORFOLOGICA + REGIAO:FAIXA.ETA
RIA
##
##                         Df Deviance    AIC
## + TONICIDADE           1   9137.3 9185.3
## <none>                     9154.9 9200.9
## - REGIAO:FAIXA.ETARIA  2   9201.0 9243.0
## - SEXO.GENERO          1   9264.4 9308.4
## - CLASSE.MORFOLOGICA   5   9286.6 9322.6
## - POSICAO.R            1   9289.8 9333.8
## - CONT.FON.SEG         3   9321.8 9361.8
## - CONT.FON.PREC        6   9439.8 9473.8
## - INDICE.SOCIO         1   9520.7 9564.7
##
## Step:  AIC=9185.26
## VD ~ CONT.FON.PREC + INDICE.SOCIO + REGIAO + FAIXA.ETARIA + CONT.FO
N.SEG +
##     SEXO.GENERO + POSICAO.R + CLASSE.MORFOLOGICA + TONICIDADE +
##     REGIAO:FAIXA.ETARIA
##
##                         Df Deviance    AIC
## <none>                     9137.3 9185.3
## - TONICIDADE           1   9154.9 9200.9
## - REGIAO:FAIXA.ETARIA  2   9183.0 9227.0
```

```
## - POSICAO.R            1   9222.9 9268.9
## - SEXO.GENERO          1   9247.3 9293.3
## - CLASSE.MORFOLOGICA   5   9282.9 9320.9
## - CONT.FON.SEG         3   9304.3 9346.3
## - CONT.FON.PREC        6   9369.8 9405.8
## - INDICE.SOCIO         1   9508.3 9554.3
```

E veja o resultado de `m.both`.

```
m.both

##
## Call:  glm(formula = VD ~ CONT.FON.PREC + INDICE.SOCIO + REGIAO + F
AIXA.ETARIA +
##     CONT.FON.SEG + SEXO.GENERO + POSICAO.R + CLASSE.MORFOLOGICA +
##     TONICIDADE + REGIAO:FAIXA.ETARIA, family = binomial, data = dad
os)
##
## Coefficients:
##                       (Intercept)                    CONT.FON.PRECe
##                            1.0941                            0.3175
##                   CONT.FON.PREC3                    CONT.FON.PRECa
##                            0.5276                            0.5040
##                   CONT.FON.PREC0                    CONT.FON.PRECo
##                            0.4788                           -0.3160
##                   CONT.FON.PRECu                      INDICE.SOCIO
##                           -0.9571                           -0.9014
##                 REGIAOperiferica                    FAIXA.ETARIA2a
##                            1.2861                           -0.1453
##                   FAIXA.ETARIA3a               CONT.FON.SEGcoronal
##                           -0.4594                            0.6786
##               CONT.FON.SEGdorsal                CONT.FON.SEGlabial
##                           -0.2168                            0.2368
##             SEXO.GENEROmasculino                   POSICAO.Rmedial
##                            0.5481                           -0.8282
##      CLASSE.MORFOLOGICAadverbio      CLASSE.MORFOLOGICAconj.prep
##                            1.0456                            0.4792
##      CLASSE.MORFOLOGICAmorf.inf    CLASSE.MORFOLOGICAsubstantivo
##                           -1.1081                            0.1249
##         CLASSE.MORFOLOGICAverbo                 TONICIDADEtonica
##                            0.4779                            0.2954
## REGIAOperiferica:FAIXA.ETARIA2a  REGIAOperiferica:FAIXA.ETARIA3a
##                           -0.7167                           -0.7676
##
## Degrees of Freedom: 9225 Total (i.e. Null);  9202 Residual
## Null Deviance:       10990
## Residual Deviance: 9137  AIC: 9185
```

Todas as variáveis são novamente selecionadas e os coeficientes coincidem. Além da função `step()`, podemos aplicar a função `drop1()` também a modelos de regressão logística para verificar a significância de cada variável previsora no modelo e se alguma

deve ser descartada. Mas diferentemente do modelo de regressão linear, aqui usaremos test = "LR". Digite então `drop1(modelo, test = "LR")`.

```
drop1(modelo, test = "LR")
```

```
## Single term deletions
##
## Model:
## VD ~ SEXO.GENERO + FAIXA.ETARIA * REGIAO + INDICE.SOCIO + CONT.FON.
PREC +
##     CONT.FON.SEG + TONICIDADE + POSICAO.R + CLASSE.MORFOLOGICA
##                       Df Deviance    AIC    LRT  Pr(>Chi)
## <none>                     9137.3 9185.3
## SEXO.GENERO            1   9247.3 9293.3 110.02 < 2.2e-16 ***
## INDICE.SOCIO           1   9508.3 9554.3 371.00 < 2.2e-16 ***
## CONT.FON.PREC          6   9369.8 9405.8 232.50 < 2.2e-16 ***
## CONT.FON.SEG           3   9304.3 9346.3 167.05 < 2.2e-16 ***
## TONICIDADE             1   9154.9 9200.9  17.66 2.636e-05 ***
## POSICAO.R              1   9222.9 9268.9  85.63 < 2.2e-16 ***
## CLASSE.MORFOLOGICA     5   9282.9 9320.9 145.60 < 2.2e-16 ***
## FAIXA.ETARIA:REGIAO    2   9183.0 9227.0  45.69 1.198e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

O resultado de `drop1()` também informa que todas as variáveis previsoras do modelo são significativas e devem ser mantidas.

Ao incluir muitas variáveis previsoras num modelo, vale a pena verificar se não estamos "inflacionando" os parâmetros de previsão. Isso é chamado de "sobreajuste" (= overfitting), a inclusão de mais variáveis do que necessário, algo que viola o Princípio da Navalha de Occam. Isso pode ser testado por meio da função `validate()`, do pacote rms (que já carregamos acima). A função `validate()` precisa de um modelo criado com a função `lrm()` que, além da fórmula e do conjunto de dados, recebe mais dois argumentos: `x = T` e `y = T`. Rode esta linha de comando, que já está pronta.

```
mod.lrm <- lrm (VD ~
                  SEXO.GENERO +
                  FAIXA.ETARIA * REGIAO +
                  INDICE.SOCIO +
                  CONT.FON.PREC +
                  CONT.FON.SEG +
                  TONICIDADE +
                  POSICAO.R +
                  CLASSE.MORFOLOGICA,
                data = dados, x = T, y = T)
```

Podemos agora aplicar a função `validate()` ao modelo criado acima. O primeiro argumento da função é o modelo de regressão logística; o segundo argumento, B, é o número de vezes que o modelo será testado por meio de "bootstrapping"; o terceiro argumento é o modo de seleção. Vamos fazer isso 200 vezes, com modo de seleção "de trás para frente". Digite então `validate(mod.lrm, B = 200, bw = T)`. (Não se assuste se demorar. Enquanto houver um ícone redondo vermelho na parte superior da janela Console, aguarde!)

```
validate(mod.lrm, B = 200, bw = T)

##
##      Backwards Step-down - Original Model
##
## No Factors Deleted
##
## Factors in Final Model
##
##  [1] SEXO.GENERO           FAIXA.ETARIA          REGIAO
##  [4] INDICE.SOCIO          CONT.FON.PREC         CONT.FON.SEG
##  [7] TONICIDADE            POSICAO.R             CLASSE.MORFOLOGICA
## [10] FAIXA.ETARIA * REGIAO

##           index.orig training    test optimism index.corrected   n
## Dxy           0.5487   0.5521  0.5458   0.0063          0.5423 200
## R2            0.2617   0.2649  0.2589   0.0059          0.2558 200
## Intercept     0.0000   0.0000 -0.0098   0.0098         -0.0098 200
## Slope         1.0000   1.0000  0.9823   0.0177          0.9823 200
## Emax          0.0000   0.0000  0.0056   0.0056          0.0056 200
## D             0.2010   0.2037  0.1987   0.0050          0.1961 200
## U            -0.0002  -0.0002  0.0000  -0.0003          0.0000 200
## Q             0.2013   0.2039  0.1986   0.0052          0.1960 200
## B             0.1630   0.1623  0.1635  -0.0012          0.1642 200
## g             1.3331   1.3471  1.3234   0.0237          1.3095 200
## gp            0.2213   0.2223  0.2200   0.0023          0.2190 200
##
## Factors Retained in Backwards Elimination
##
##  SEXO.GENERO FAIXA.ETARIA REGIAO INDICE.SOCIO CONT.FON.PREC
##  *           *            *      *            *
##  *           *            *      *            *
##  *           *            *      *            *
##  *           *            *      *            *
##  *           *            *      *            *
##  *           *            *      *            *
##  *           *            *      *            *
##  *           *            *      *            *
##  *           *            *      *            *
```

```
## [...]
##  CONT.FON.SEG TONICIDADE POSICAO.R CLASSE.MORFOLOGICA
##  *            *          *         *
##  *            *          *         *
##  *            *          *         *
##  *            *          *         *
##  *            *          *         *
##  *            *          *         *
##  *            *          *         *
##  *            *          *         *
##  *            *          *         *
## [...]
##  FAIXA.ETARIA * REGIAO
##  *
##  *
##  *
##  *
##  *
##  *
##  *
##  *
##  *
## [...]
##
## Frequencies of Numbers of Factors Retained
##
##   9  10
##   1 199
```

N.B.: Resultado aqui abreviado.

"Bootstrapping", em informática, refere-se a um processo autossustentável que se implementa sem ajuda externa. O que a função `validate()` faz é selecionar randomicamente subamostras a partir do conjunto completo de dados, realizar o mesmo teste estatístico repetidas vezes (acima, fizemos 200), e verificar se os mesmos resultados se mantêm. Se há um número demasiado de variáveis previsoras para aquele número de dados, a função `validate()` vai acusar "otimismo" nos resultados. A coluna que nos interessa aqui é justamente "optimism", que mostra a diferença entre o treinamento do modelo e as medidas estatísticas calculadas. Um modelo válido tem valores abaixo de 0,05 na coluna "optimism".

Em seguida, o resultado mostra as variáveis que foram selecionadas em cada um dos 200 testes, por meio dos asteriscos (às vezes com linhas omitidas). Ao final, o teste informa quantas variáveis foram selecionadas quantas vezes. Neste conjunto de dados,

podemos ficar tranquilos que a inclusão de 10 variáveis previsoras não sobreajusta o modelo, pois o número de vezes que as 10 variáveis previsoras foram mantidas é maior do que o número de vezes que alguma foi eliminada. Caso isso não ocorresse, a solução seria buscar mais dados ou diminuir o número de variáveis previsoras do modelo. Aquelas selecionadas por último em step forward são as melhores candidatas à exclusão. Vale mencionar que a função `validate()` também pode ser aplicada a modelos lineares (ver Levshina, 2015, cap.7).

Sua análise de regressão logística ainda não terminou! Assim como nos modelos lineares, é importante checar se os pressupostos da regressão logística foram atendidos. Um deles é que a relação entre as estimativas e as variáveis previsoras numéricas é linear. Aqui, temos apenas uma variável previsora numérica, INDICE.SOCIO. Para fazer este teste, vamos aplicar a função `crPlot()` do pacote car, que toma um modelo e a variável previsora como argumentos. Contudo, infelizmente, a função `crPlot()` não aceita modelos com interação. Façamos então um novo modelo, chamado `modelo2`, que contém a fórmula que vimos usando, mas sem interação.

```
modelo2 <- glm(VD ~
                 SEXO.GENERO +
                 FAIXA.ETARIA +
                 REGIAO +
                 INDICE.SOCIO +
                 CONT.FON.PREC +
                 CONT.FON.SEG +
                 TONICIDADE +
                 POSICAO.R +
                 CLASSE.MORFOLOGICA,
               data = dados, family = binomial)
```

Aplique agora a função `crPlot()` a `modelo2`, com o segundo argumento var = "INDICE.SOCIO".

```
crPlot(modelo2, var = "INDICE.SOCIO")
```

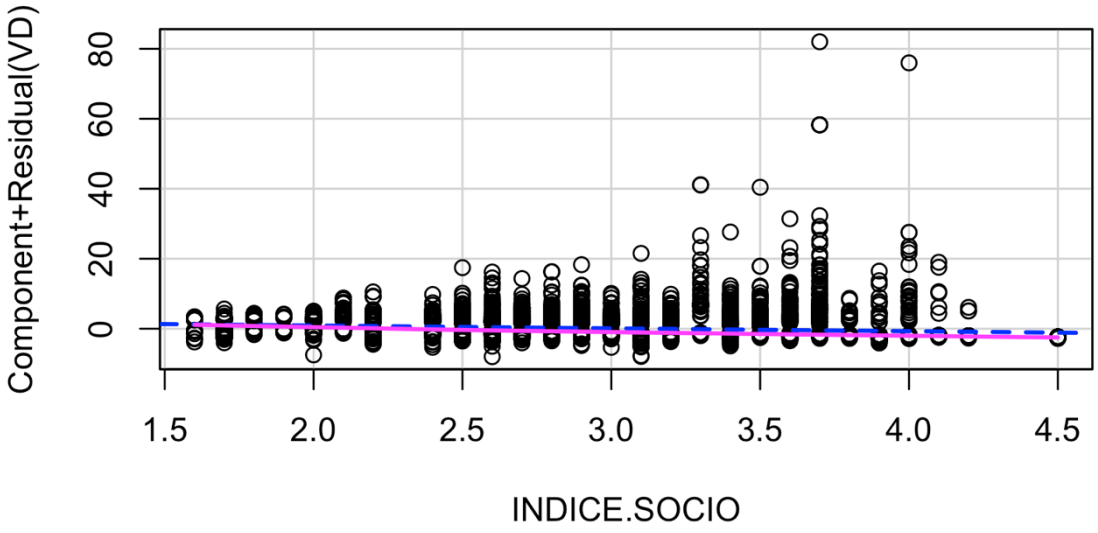

Figura 15.1: Plot dos valores previstos e observados de INDICE.SOCIO no modelo2. Fonte: própria.

Vemos que a linha pontilhada, de valores previstos, e a linha contínua, de valores observados, são praticamente coincidentes. Se não fossem, deveríamos considerar a possibilidade de retirar valores atípicos da distribuição (ver Lição 13).

Outro pressuposto de modelos de regressão logística é a inexistência de multicolinearidade, que pode ser testada pela função `vif()`. Digite então `car::vif(modelo)`. (Aqui, usamos a notação pacote::funcao pois a função `vif()` também existe no pacote rms, mas queremos usar a do pacote car.)

```
car::vif(modelo)

##                          GVIF Df GVIF^(1/(2*Df))
## SEXO.GENERO          1.032890  1        1.016312
## FAIXA.ETARIA         6.564032  2        1.600636
## REGIAO               2.806091  1        1.675139
## INDICE.SOCIO         1.076890  1        1.037733
## CONT.FON.PREC        3.985509  6        1.122123
## CONT.FON.SEG         2.028554  3        1.125117
## TONICIDADE           1.828523  1        1.352229
## POSICAO.R            2.095435  1        1.447562
## CLASSE.MORFOLOGICA   3.408764  5        1.130472
## FAIXA.ETARIA:REGIAO 10.667029  2        1.807219
```

Vemos aí que a maior parte dos valores estão bem abaixo de 5, o que indicia não colinearidade entre as variáveis do modelo. Os valores mais altos são de FAIXA.ETARIA e a interação FAIXA.ETARIA:REGIAO; aqui, é esperado que haja falta de colinearidade, já que a interação envolve a variável. A terceira coluna, de GVIF-ajustado, leva isso em conta e corrige a medida.

Por fim, deve-se checar se as observações do conjunto de dados são independentes umas das outras. Como já se notou na Lição 13, isso raramente é o caso de análises linguísticas, de modo que é sempre recomendável realizar *análises de efeitos mistos*, com a inclusão de efeitos aleatórios. Nos dados de /r/ em coda, as variáveis PARTICIPANTE e ITEM.LEXICAL representam efeitos aleatórios.

Vimos que modelos de efeitos mistos de regressão linear são criados por meio da função `lmer()`, por oposição à função `lm()`. Em modelos de efeitos mistos de regressão logística, a função é... `glmer()`! Lembre-se que efeitos aleatórios entram na fórmula com a notação (1|varaleatoria). Rode então a linha de comando a seguir. Uma nota: não estranhe se esse modelo demorar 3-4 minutos para rodar!

```
mod.glmer <- glmer(VD ~
                     SEXO.GENERO +
                     FAIXA.ETARIA * REGIAO +
                     INDICE.SOCIO +
                     CONT.FON.PREC +
                     CONT.FON.SEG +
                     TONICIDADE +
                     POSICAO.R +
                     CLASSE.MORFOLOGICA +
                     (1|PARTICIPANTE) +
                     (1|ITEM.LEXICAL),
                   data = dados, family = binomial)
```

Visualize agora o resultado com `summary(mod.glmer)`.

```
summary(mod.glmer)

## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula:
## VD ~ SEXO.GENERO + FAIXA.ETARIA * REGIAO + INDICE.SOCIO + CONT.FON.
PREC +
##     CONT.FON.SEG + TONICIDADE + POSICAO.R + CLASSE.MORFOLOGICA +
##     (1 | PARTICIPANTE) + (1 | ITEM.LEXICAL)
##    Data: dados
```

```
##
##      AIC      BIC   logLik deviance df.resid
##   7507.5   7692.9  -3727.8   7455.5     9200
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -7.8597 -0.4393 -0.2085  0.3127 11.0416
##
## Random effects:
##  Groups       Name        Variance Std.Dev.
##  ITEM.LEXICAL (Intercept) 0.4814   0.6939
##  PARTICIPANTE (Intercept) 2.0592   1.4350
## Number of obs: 9226, groups:  ITEM.LEXICAL, 1151; PARTICIPANTE, 118
##
## Fixed effects:
##                                 Estimate Std. Error z value Pr(>|z
|)
## (Intercept)                      0.439246   0.925710   0.474  0.635
15
## SEXO.GENEROmasculino             0.790969   0.275991   2.866  0.004
16
## FAIXA.ETARIA2a                   0.336203   0.505514   0.665  0.506
00
## FAIXA.ETARIA3a                  -0.126504   0.497683  -0.254  0.799
35
## REGIAOperiferica                 2.005981   0.496544   4.040 5.35e-
05
## INDICE.SOCIO                    -1.117505   0.237855  -4.698 2.62e-
06
## CONT.FON.PRECe                   0.284838   0.273786   1.040  0.298
17
## CONT.FON.PREC3                   0.589093   0.320176   1.840  0.065
78
## CONT.FON.PRECa                   0.661105   0.275211   2.402  0.016
30
## CONT.FON.PREC0                   0.641808   0.312345   2.055  0.039
90
## CONT.FON.PRECo                  -0.331198   0.280419  -1.181  0.237
57
## CONT.FON.PRECu                  -1.418130   0.338658  -4.188 2.82e-
05
## CONT.FON.SEGcoronal              0.659775   0.161881   4.076 4.59e-
05
## CONT.FON.SEGdorsal               0.008341   0.189613   0.044  0.964
91
## CONT.FON.SEGlabial               0.315251   0.177475   1.776  0.075
68
## TONICIDADEtonica                 0.558847   0.128725   4.341 1.42e-
05
## POSICAO.Rmedial                 -0.689526   0.162930  -4.232 2.32e-
05
## CLASSE.MORFOLOGICAadverbio       0.850362   0.486957   1.746  0.080
76
## CLASSE.MORFOLOGICAconj.prep      0.719349   0.317131   2.268  0.023
```

```
31
## CLASSE.MORFOLOGICAmorf.inf      -0.792194   0.262457  -3.018  0.002
54
## CLASSE.MORFOLOGICAsubstantivo    0.247449   0.134095   1.845  0.064
99
## CLASSE.MORFOLOGICAverbo          0.800943   0.161455   4.961 7.02e-
07
## FAIXA.ETARIA2a:REGIAOperiferica -1.128955   0.680637  -1.659  0.097
18
## FAIXA.ETARIA3a:REGIAOperiferica -1.258274   0.681851  -1.845  0.064
98
##
## (Intercept)
## SEXO.GENEROmasculino            **
## FAIXA.ETARIA2a
## FAIXA.ETARIA3a
## REGIAOperiferica                ***
## INDICE.SOCIO                    ***
## CONT.FON.PRECe
## CONT.FON.PREC3                  .
## CONT.FON.PRECa                  *
## CONT.FON.PREC0                  *
## CONT.FON.PRECo
## CONT.FON.PRECu                  ***
## CONT.FON.SEGcoronal             ***
## CONT.FON.SEGdorsal
## CONT.FON.SEGlabial              .
## TONICIDADEtonica                ***
## POSICAO.Rmedial                 ***
## CLASSE.MORFOLOGICAadverbio      .
## CLASSE.MORFOLOGICAconj.prep     *
## CLASSE.MORFOLOGICAmorf.inf      **
## CLASSE.MORFOLOGICAsubstantivo   .
## CLASSE.MORFOLOGICAverbo         ***
## FAIXA.ETARIA2a:REGIAOperiferica .
## FAIXA.ETARIA3a:REGIAOperiferica .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.0387909 (tol = 0.002, c
omponent 1)
```

No modelo de efeitos mistos, quais variáveis deixam de ser significativamente correlacionadas com VD?

- CONT.FON.PREC e CONT.FON.SEG
- TONICIDADE e POSICAO.R
- CLASSE.MORFOLOGICA e CONT.FON.PREC
- FAIXA.ETARIA e a interação com REGIAO

Como já argumentado várias vezes neste curso, a interpretação de resultados numéricos é sempre mais fácil por meio de gráficos. Faça um gráfico de efeitos com `plot(allEffects(mod.glmer), type = "response")`.

```
plot(allEffects(mod.glmer), type = "response")
```

N.B.: Resultado aqui omitido.

Também se pode adicionar o argumento ask = T para selecionar manualmente quais efeitos você quer visualizar. A partir da linha de comando acima, adicione o novo argumento. Clique sobre a variável ou variáveis de interesse e clique sobre "cancelar" quando não quiser incluir mais nenhuma.

```
plot(allEffects(mod.glmer), type = "response", ask = T)
```

N.B.: Resultado aqui omitido.

Deixei disponível no Anexo D uma sugestão de roteiro de análise para variáveis nominais binárias, tomando o conjunto de dados da realização do /r/ em coda como exemplo. Assim como no roteiro da Lição 13, a ideia é sistematizar os passos de uma análise que ficaram espalhados por várias lições deste curso. Para consultar como os resultados de modelos de regressão logística podem ser reportados, reveja o final da Lição 13, sobretudo a Figura 13.5.

E isso conclui a Lição 15 e este curso. Mas há muito mais a se aprender! Espero que este curso tenha sido apenas o ponto de partida!

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, vamos trabalhar sobre os dados da realização de /r/ em lojas de Departamento em Nova Iorque, de Labov. Primeiro, carregue os dados da planilha LabovDS.csv. Após carregar o dataframe, cheque sua estrutura, como de praxe. Exclua os dados de d da variável r. Certifique-se de que o nível de referência é r1 – para que os

resultados de regressão logística sejam lidos em referência ao apagamento de /r/ (r0). Reorganize os níveis das variáveis store, word e emphasis em ordem alfabética.

Essa planilha contém os dados da realização de /r/ (r0 – apagado; r1 – realizado) por parte de funcionários de três lojas de departamento em Nova Iorque. As variáveis previsoras são store (Klein, Macy's, Saks), emphasis (casual ou emphasic) e word (fourth ou floor). Verifique se há interações entre essas três variáveis em modelos simples de regressão logística, que incluem apenas uma interação.

1. Entre qual par de variáveis há interação?
    a. nenhum
    b. emphasis e word
    c. store e emphasis
    d. store e word
2. Crie um modelo com todas as variáveis. Inclua a interação, se houver. Qual variável não apresenta correlação significativa com a pronúncia de /r/?
    a. emphasis
    b. store
    c. word
    d. todas têm correlação significativa
3. Qual é o índice C deste modelo?
4. Qual é o valor de R^2 deste modelo?
5. Este modelo, como um todo, é significativo? Justifique sua resposta.
6. Os funcionários de qual loja mais favorecem o apagamento de /r/?
    a. Klein
    b. Macy's
    c. Saks
7. Qual item lexical mais favorece o apagamento de /r/?
    a. fourth
    b. floor

8. Calcule a probabilidade (de 0% a 100%) de se realizar o apagamento de /r/ na palavra fourth. Reporte o resultado em número decimal.
9. Acima se testou se há interações. Teste se há colinearidade entre as variáveis previsoras.
    a. não há
    b. há entre emphasis e word
    c. há entre store e emphasis
    d. há entre store e word
10. Neste conjunto de dados, que tipo de variável é word?
    a. efeito fixo
    b. efeito aleatório