

Análise Quantitativa de Dados em Linguística

Regressão Logística

Ronaldo Lima Jr.

`ronaldojr@letras.ufc.br`

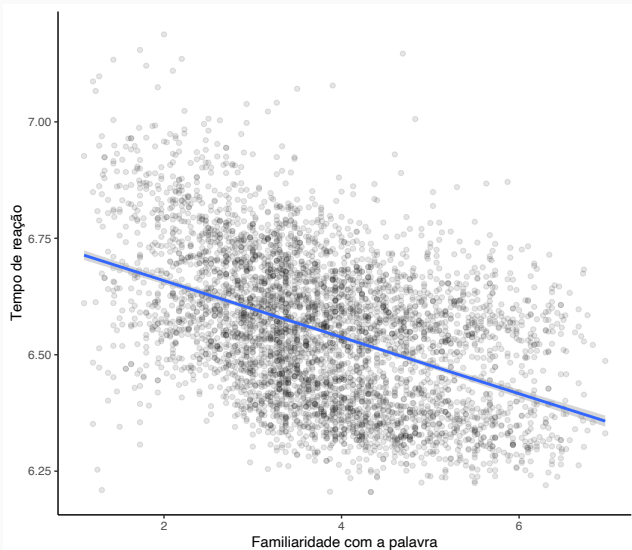
`ronaldolimajr.github.io`

Universidade Federal do Ceará

1. Regressão Linear (revisão)
2. Regressão Logística
3. Modelo 1
4. Modelo 2

Regressão Linear (revisão)

Regressão Linear

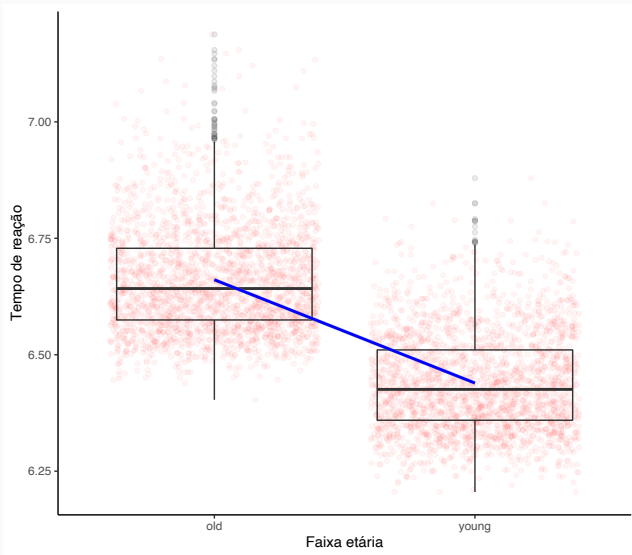


```
1 | RTlexdec ~ Familiarity
```

→ Variável resposta contínua

→ Variável preditora contínua

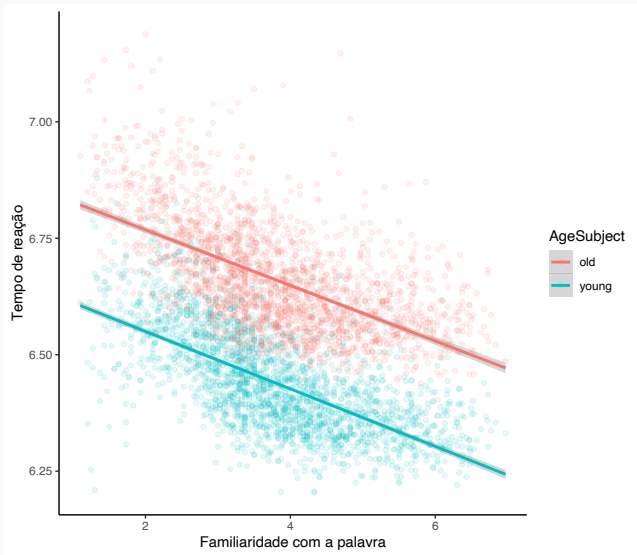
Regressão Linear



```
1 | RTlexdec ~ AgeSubject
```

- Variável resposta contínua
- Variável preditora categórica

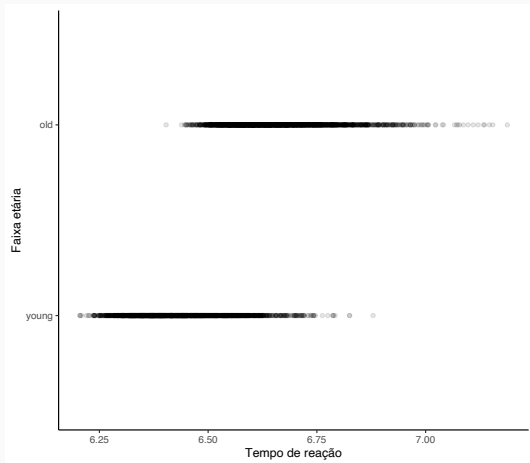
Regressão Linear



1 | `RTlexdec ~ Familiarity + AgeSubject`

- Variável resposta contínua
- Variáveis preditoras contínuas e categóricas

Regressão Linear



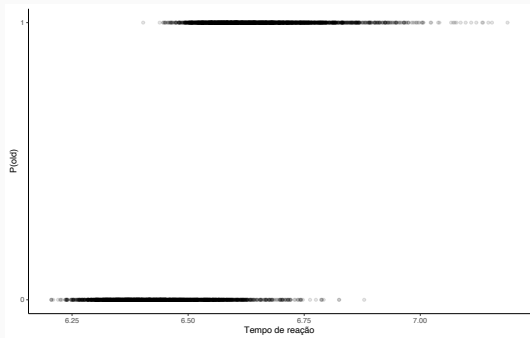
E se a variável preditora for binária? (yes/no, 0/1, good/bad, natural/unnatural, etc.)

Ex.: modelar faixa etária em função do tempo de reação

i.e.: prever se um participante é jovem ou velho com base no seu tempo de reação

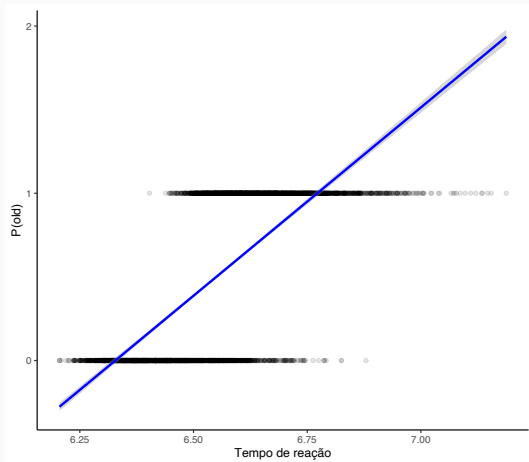
Tarefas de classificação como essa são muito comuns em *machine learning* (e é o que o Goldvarb faz por trás)

Regressão Linear



Nesses casos, não há uma unidade de medida no eixo y , então precisamos modelar/prever a **probabilidade** de um participante ser velho ou jovem diante de um tempo de reação.

Regressão Linear



Podemos simplesmente ajustar uma linha a esses dados?

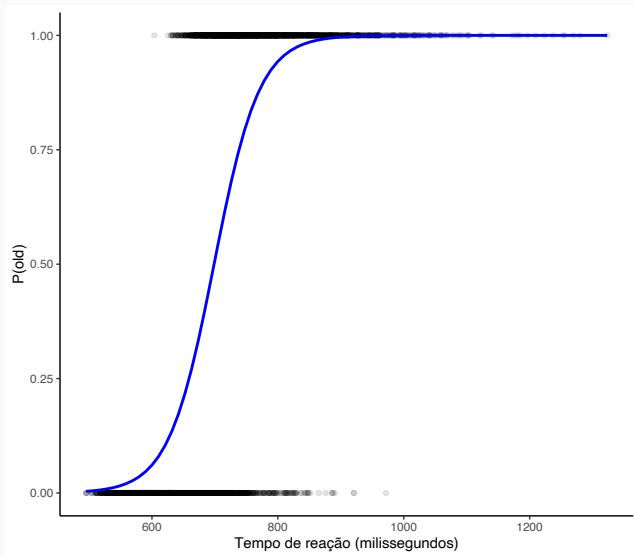
```
1 | lm(AgeSubject ~ RTLexdec)
```

- Claramente, **não**!
- A maioria dos dados está longe da reta
- Probabilidades vão de 0 a 1, e a reta estima probabilidades acima de 1 e abaixo de 0 (impossíveis)

O que fazer?

Regressão Logística

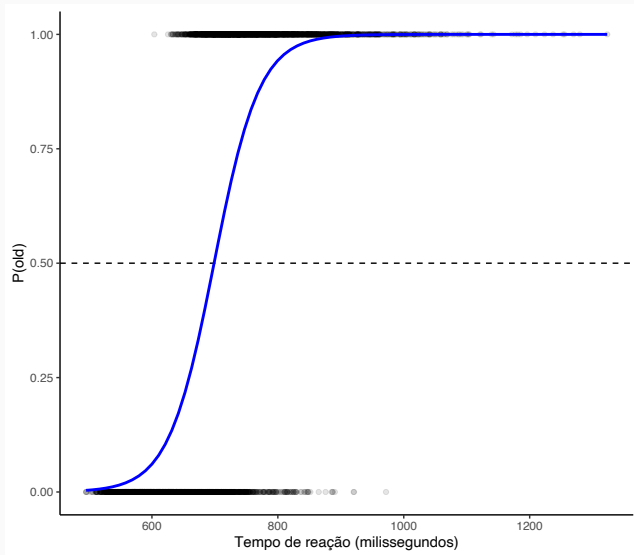
Regressão Logística



Utilizamos uma linha curva!

→ Neste caso, uma **curva logística** (ou sigmoide)

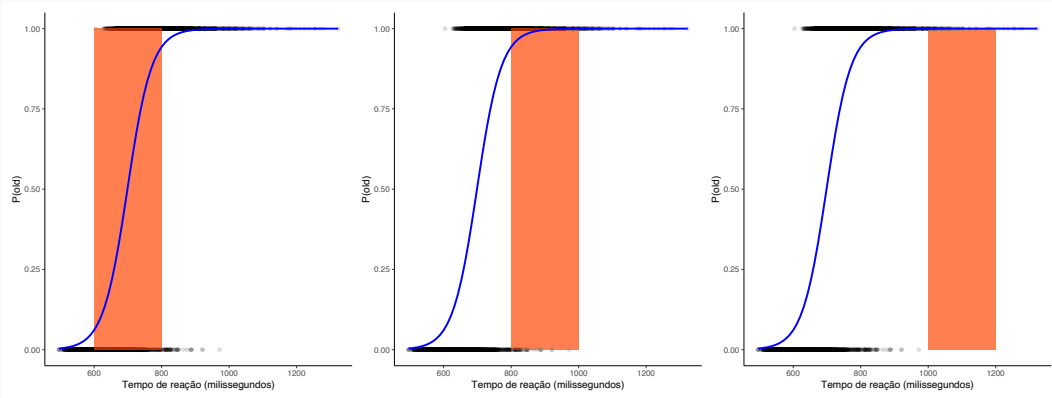
Regressão Logística



Utilizamos uma linha curva!

- Neste caso, uma **curva logística** (ou sigmoide)
- Porém, não podemos modelar a probabilidade em função das variáveis preditoras (neste caso $P(\text{old})$ em função do tempo de reação) porque a probabilidade não é constante (não é uma linha reta).

Mudanças de 200 milissegundos no tempo de reação:



Solução:

- Utilizar uma medida relacionada a probabilidades, porém constante (linear), simétrica e facilmente reconvertida para probabilidades
- *log-odds* (logaritmo das chances)

Odds (chances)

- As chances (odds) de algo ocorrer são a razão (divisão) da quantidade favorável pela quantidade desfavorável:
 - chances de um dado (justo) cair em um número específico = $1:5 = 0,2$
 - chances de uma moeda (justa) cair cara = $1:1 = 1$
 - Labov observou 195 apagamentos de /r/ e 21 produções de /r/ da loja Klein → chances de apagamento da loja Klein = $195:21 = 9,3$
- As chances vão de 0 a ∞ , com ponto de equilíbrio em 1:
 - chances entre 0 e 1: desfavorecem a ocorrência (dado)
 - chances de 1: ponto neutro/equilíbrio (moeda)
 - Chances maiores que 1: favorecem a ocorrência (apagamento de /r/ da Klein)

Problema: a escala das chances não é simétrica

$$0 \longleftrightarrow 1 \longleftrightarrow \infty$$

Probabilidade

- A probabilidade de algo ocorrer é a razão (divisão) da quantidade favorável pela quantidade total:
 - probabilidade de um dado (justo) cair em um número específico = $1/6 = 0,1666 \approx 17\%$
 - probabilidade de uma moeda (justa) cair cara = $1/2 = 0,5 = 50\%$
 - Labov observou 195 apagamentos de /r/ e 21 produções de /r/ da loja Klein → probabilidade de apagamento da loja Klein = $195/216 = 0,9027778 \approx 90\%$
- Probabilidade vai de 0 a 1, com ponto de equilíbrio em 0,5 (a escala é simétrica):
 - probabilidades entre 0 e 0,5 ($< 50\%$): desfavorecem a ocorrência (dado)
 - probabilidade de 0,5 (50%): ponto neutro/equilíbrio (moeda)
 - probabilidades maiores que 0,5 ($> 50\%$): favorecem a ocorrência (apagamento de /r/ da Klein)

Problema: o ponto de equilíbrio em 0,5 nem sempre é intuitivo; e, a mudança de probabilidade na curva logística (sigmoide) que utilizaremos não é constante

- Ao tirarmos o logaritmo natural (\log) das chances (odds), resolvemos o problema da assimetria da escala das chances, pois os valores de 0 a 1 ficam negativos; e resolvemos a não constância da probabilidade na curva logística, pois os valores em log-odds serão constantes
- A escala em log-odds vai de $-\infty$ a $+\infty$
- É simples fazer uma conversão entre qualquer uma dessas três medidas (probabilidade, chances/odds, log-odds)

- probabilidade \leftrightarrow odds
 - $p = odds / 1 + odds$
 - $odds = p / 1 - p$
- odds \leftrightarrow log-odds
 - $\text{log-odds} = \log(odds)$
 - $odds = \exp(\text{log-odds})$
- probabilidade \leftrightarrow log-odds
 - $\text{logodds} = \log(p / 1 - p)$
 - $p = 1 / 1 + \exp(-\text{logodds})$

```
1 | arm::invlogit()  
2 | ilogit = function(x) {  
3 |   1/(1+exp(-x))}
```

Conversões

Probabilidade	Chances (odds)	Log-odds
0,001	0,001	-6,91
0,01	0,01	-4,6
0,05	0,51	-2,94
0,1	0,11	-2,2
0,25	0,33	-1,1
0,5	1	0
0,75	3	1,1
0,9	9	2,2
0,95	19	2,94
0,99	99	4,6
0,999	999	6,91

- Obs. sobre log-odds:
 - valores negativos = baixa probabilidade (desfavorece)
 - valores positivos = alta probabilidade (favorece)
 - log-odd de 2 \approx 90% (ou 10%) de probabilidade
 - log-odd de 4 \approx 99% (ou 1%) de probabilidade

- Modelos Lineares e Modelos Logísticos são parte de uma mesma família

Generalized Linear Models (GLMs):

Linear, Logistic, Ordinal, Poisson, Multinomial, etc.

- Consequentemente, o output de uma regressão linear e de uma regressão logística é muito semelhante
- Principais diferenças:
 - coeficientes são dados em log-odds (conversíveis para probabilidade)
 - não há R^2 , pois não há resíduos

Modelo 1

Exemplo 1 (modelo com 1 variável preditora contínua)

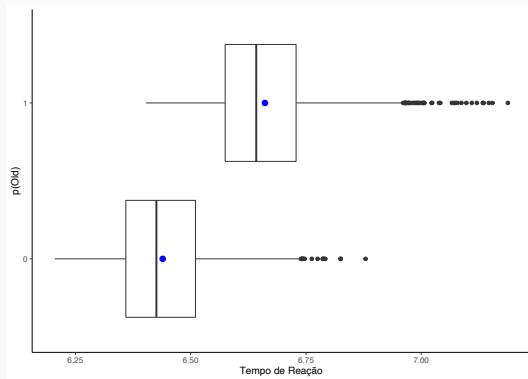
Dados 'english' do pacote `languageR`

- Com modelo de regressão linear buscamos prever tempo de reação a partir da faixa etária (`RTlexdec` em função de `AgeSubject`)
- Podemos agora prever a faixa etária a partir do tempo de reação (`AgeSubject` em função de `RTlexdec`)?

Modelo de Regressão Logística 1

Contrast Coding:

```
1 english = english %>%
2   dplyr::select(RTlexdec, AgeSubject) %>%
3   mutate(Old = as.factor(
4     ifelse(AgeSubject == "old", 1, 0)
5   ))
6
7   RTlexdec AgeSubject Old
8 1  6.543754      young  0
9 2  6.397596      young  0
10 3  6.304942      young  0
11 4  6.424221      young  0
12 5  6.450597      young  0
```



Modelo de Regressão Logística 1

```
1 | model.eng = glm(Old ~ RTlexdec, data = english, family = binomial())
2 | summary(model.eng)

1 | Deviance Residuals:
2 |      Min       1Q   Median       3Q      Max
3 | -3.6178  -0.4229  -0.0235   0.5312   2.3951

4 | Coefficients:
5 |             Estimate Std. Error z value Pr(>|z|)
6 | (Intercept) -128.6308    3.8008  -33.84  <2e-16 ***
7 | RTlexdec     19.6497     0.5804   33.86  <2e-16 ***
8 | ---
9 | Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

10 | (Dispersion parameter for binomial family taken to be 1)

11 |      Null deviance: 6332.6  on 4567  degrees of freedom
12 | Residual deviance: 3145.8  on 4566  degrees of freedom
13 | AIC: 3149.8
```

Modelo de Regressão Logística 1

Interpretação dos coeficientes:

```
1 | Coefficients:
2 |             Estimate Std. Error z value Pr(>|z|)
3 | (Intercept) -128.6308    3.8008  -33.84  <2e-16 ***
4 | RTlexdec     19.6497     0.5804   33.86  <2e-16 ***
```

- *Intercept*: probabilidade de ser Old quando tempo de reação = 0
- -129 log-odds (negativo, valores de referência da tabela)
- *Slope*: quanto o *intercept* muda (neste caso aumenta) em log-odds para cada unidade de mudança em RTlexdec
- 20 log-odds a mais para cada unidade de aumento em tempo de reação (não podemos transformar este *slope* em probabilidade porque a probabilidade da curva logística não é constante)

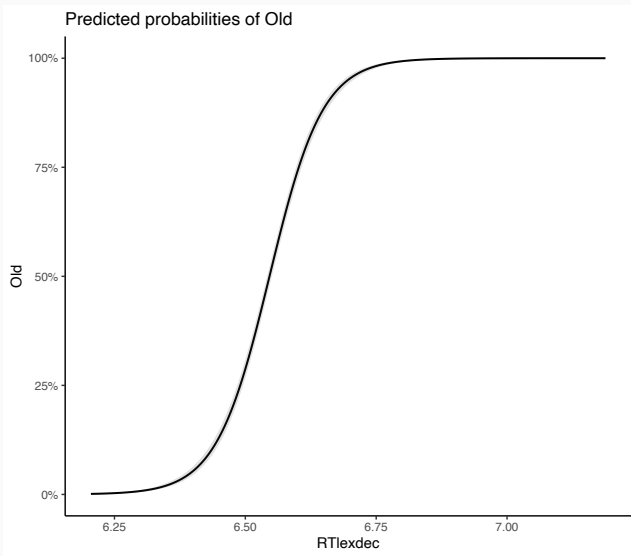
Modelo de Regressão Logística 1

Prever probabilidades de 'Old':

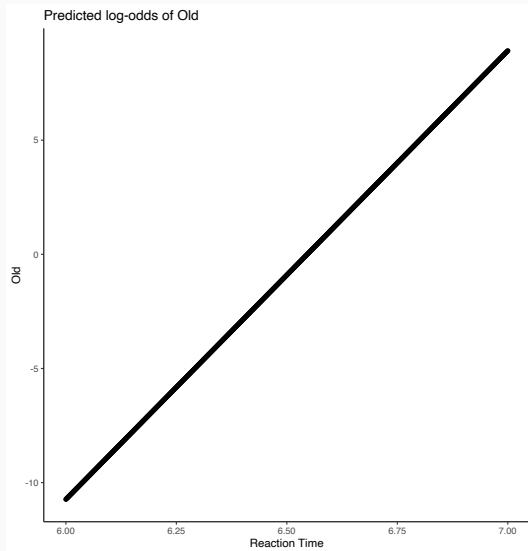
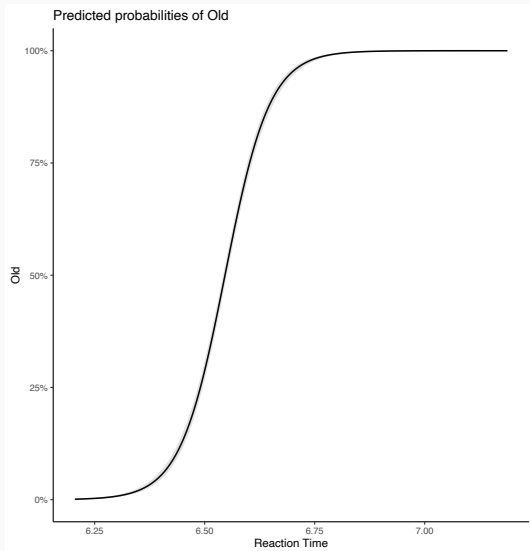
```

1  invlogit(-128.6308 + 19.6497)
2  4.678532e-48
3
3  invlogit(-128.6308 + (2*19.6497))
4  1.599064e-39
5
5  invlogit(-128.6308 + (6.25*19.6497))
6  0.002958308
7
7  invlogit(-128.6308 + (6.5*19.6497))
8  0.2874605
9
9  invlogit(-128.6308 + (6.75*19.6497))
10 0.9820962
11
11 invlogit(-128.6308 + (7*19.6497))
12 0.9998659

```



Modelo de Regressão Logística 1



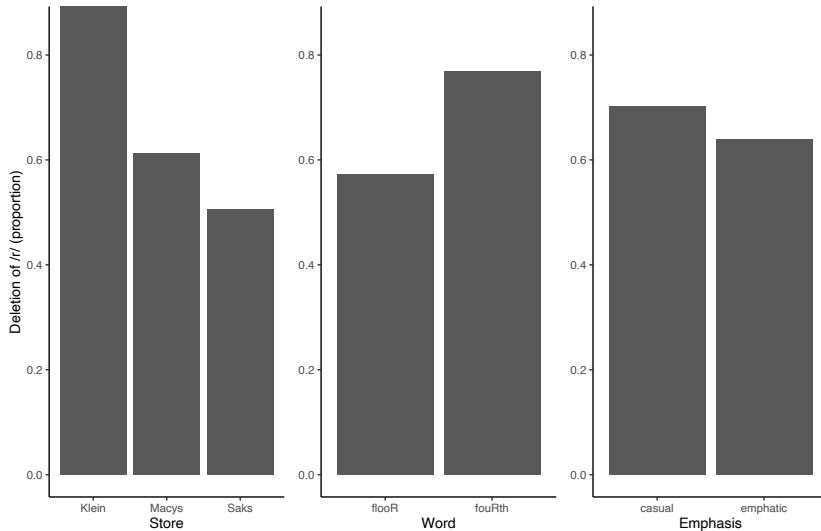
Modelo 2

Exemplo 2 (modelo com variáveis preditoras categóricas)

Dados 'Labov'

- Modelar/explicar/prever apagamento do /r/ em coda em função de:
 - loja/classe socioeconômica (Klein, Macys, Saks)
 - palavra/posição do /r/ – coda medial ou final (fouRth, flooR)
 - ênfase (casual, enfático)

Modelo de Regressão Logística 2



Modelo de Regressão Logística 2

Modelo com apenas “loja” como preditor:

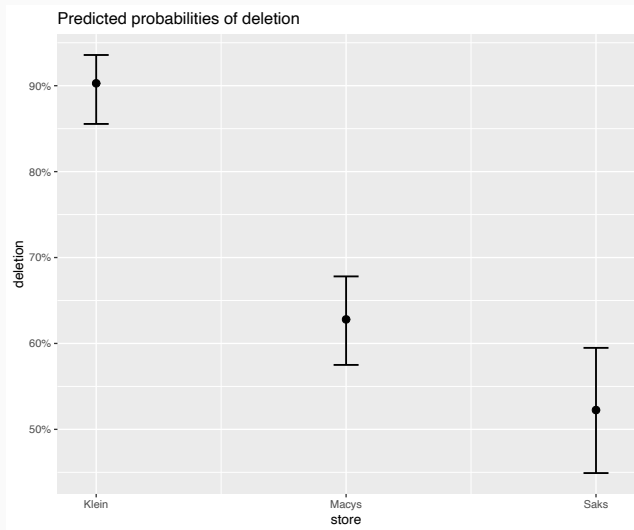
```
1 # Contrast coding para modelarmos apagamento
2 labov = labov %>%
3   mutate(deletion = if_else(r == "r0", 1, 0))
4 mLabov.store = glm(deletion ~ store,
5                     data = labov,
6                     family = "binomial")
7 summary(mLabov.store)
8
9 Coefficients:
10      (Intercept)      2.2285      0.2296      9.704 < 2e-16 ***
11 storeMacys      -1.7049      0.2559     -6.663 2.68e-11 ***
12 storeSaks       -2.1385      0.2743     -7.796 6.41e-15 ***
```

- Todos coeficientes significativos
- *Intercept* = probabilidade (em log-odds) de apagamento na loja Klein (positivo)
 $\text{arm::invlogit}(2.2285) = 0.90$
- *Slope: Macys* = mudança no intercept na Macys (negativo = probabilidade diminuir)
 $\text{invlogit}(2.2285 - 1.7049) = 0.63$
- *Slope: Saks* = mudança no intercept na Saks (negativo = probabilidade diminuir)
 $\text{invlogit}(2.2285 - 2.1385) = 0.52$

Fazemos os cálculos com os log-odds e transformamos (com `'invlogit()'`, por exemplo) apenas o resultado (e não os coeficientes)

Modelo de Regressão Logística 2

Gráfico das probabilidades previstas pelo modelo (90%, 63% e 52%) com seus intervalos de 95% de confiança:



Modelo de Regressão Logística 2

Modelo com “loja” e “palavra” como preditores:

```
1 mLabov.store.word = glm(deletion ~ store + word,
2                           data = labov,
3                           family = "binomial")
4 summary(mLabov.store.word)

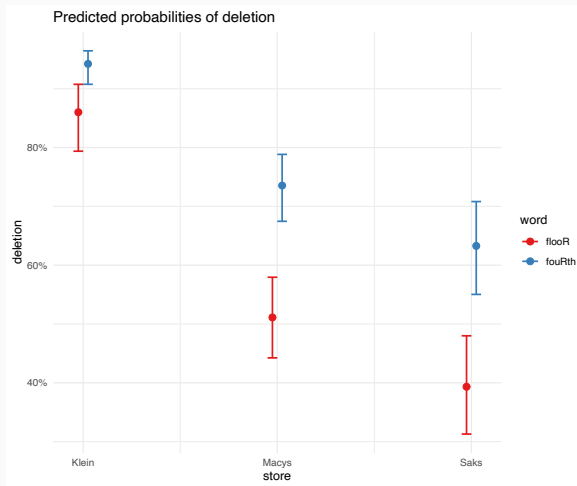
5 Coefficients:
6             Estimate Std. Error z value Pr(>|z|)
7 (Intercept)   1.8165     0.2388   7.606 2.82e-14 ***
8 storeMacys   -1.7718     0.2603  -6.806 1.00e-11 ***
9 storeSaks    -2.2500     0.2810  -8.008 1.17e-15 ***
10 wordfouRth    0.9778     0.1742   5.614 1.98e-08 ***
```

- Todos coeficientes significativos
- Os coeficientes das lojas continuam com os mesmos sinais negativos, mas com valores diferentes, pois o modelo agora tem mais informações
- O *intercept* agora representa a probabilidade (em log-odds) de apagamento da palavra “floor” na loja Klein
- Coeficiente de word:fouRth é positivo, indicando que a probabilidade de apagamento em “fourth” aumenta (em relação a “floor”)

Modelo de Regressão Logística 2

Probabilidades previstas:

```
1 pred.store.word =  
2   tibble(store = rep(levels(labov$store), times = 2),  
3           word = rep(levels(labov$word), each = 3)) %>%  
4   mutate(pred = predict(mLabov.store.word,  
5                         newdata = pred.store.word,  
6                         type = "response"))  
7 pred.store.word  
  
8 1 Klein flooR 0.860  
9 2 Macys flooR 0.511  
10 3 Saks flooR 0.393  
11 4 Klein fouRth 0.942  
12 5 Macys fouRth 0.735  
13 6 Saks fouRth 0.633
```



Modelo de Regressão Logística 2

Modelo com “loja”, “palavra” e “ênfase” como preditores:

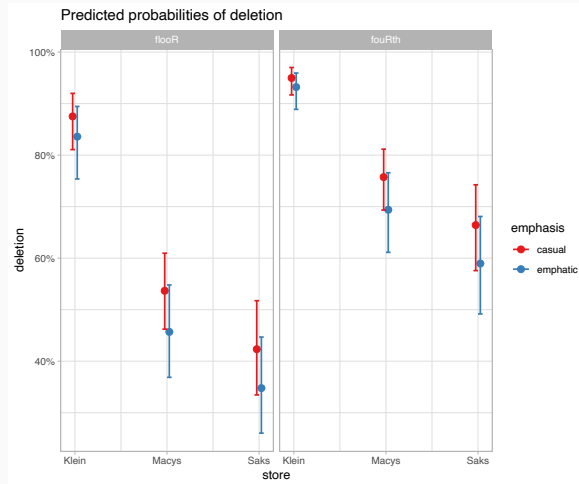
```
1 mLabov = glm(deletion ~ store + word + emphasis,  
2               data = labov,  
3               family = "binomial")  
4 summary(mLabov)  
  
5 Coefficients:  
6               Estimate Std. Error z value Pr(>|z|)  
7 (Intercept)      1.9473     0.2514   7.746 9.47e-15 ***  
8 storeMacys       -1.8004     0.2615  -6.884 5.81e-12 ***  
9 storeSaks        -2.2564     0.2817  -8.011 1.13e-15 ***  
10 wordfourth       0.9912     0.1749   5.666 1.46e-08 ***  
11 emphasisemphatic -0.3197     0.1787  -1.789  0.0736 .
```

- Coeficiente para emphasis não significativo - tamanho de efeito pequeno
- Demais coeficientes mantêm suas direções (com valores levemente diferentes)

Modelo de Regressão Logística 2

Probabilidades previstas:

```
1 pred.labov =
2   tibble(store = rep(levels(labov$store), times = 4),
3           word = rep(levels(labov$word), each = 6),
4           emphasis = rep(levels(labov$emphasis),
5                           times = 2, each = 3)) %>%
6   mutate(pred = predict(mLabov,
7                         newdata = pred.labov,
8                         type = "response"))
9
10 1 Klein flooR casual 0.875
11 2 Macys flooR casual 0.537
12 3 Saks flooR casual 0.423
13 4 Klein flooR emphatic 0.875
14 5 Macys flooR emphatic 0.537
15 6 Saks flooR emphatic 0.423
16 7 Klein fouRth casual 0.932
17 8 Macys fouRth casual 0.694
18 9 Saks fouRth casual 0.590
19 10 Klein fouRth emphatic 0.932
20 11 Macys fouRth emphatic 0.694
21 12 Saks fouRth emphatic 0.590
```



Perguntas?