Lista 1 de Lego 3

Luana Calzavara

10/04/2021

1

1.a

Em $Y_i \sim f(y|\theta, \alpha)$, temos o componente estocástico. Isto é, a representação da distribuição da nossa variável aleatória Y_i, dado θ e α. $\sim$ nos indica como está distribuída a nossa variável aleatória. Segundo Gelman and Hill(2006:13), distribuição é um conjunto de obejtos não identificados. Quando utilizamos o sinal $\sim$, queremos indicar que Y_i se distribui conforme a função seguinte, dado os parâmetros θ e α.

1.b

$f(.)$ indica a função pela qual se distribui os parâmetros que resumem a variável dependente.Ele caracteriza o componente estocástico.Parâmetros são um resumo numérico feito sobre uma população em uma inferência. No exemplo indicado, θ, em um modelo linear, é O valor esperado, em média, de Y_i. Quando a variável dependente é dicotômica, temos um π, em seu lugar, indicamento a probabilidade de $y_i = 0$ ou $y_i = 1$. α, pode ser outro parâmetro, como desvio-padrão ou variância.

Já $g(.)$ descreve o componente sistemástico. Isto é, como cada observação da var. dependente se resume em determinado θ_i. No nosso caso, θ_i é igual a função $g(.)$, para cada valor das variáveis dependentes X e seus respectivos β. Cada observação i tem seu respectivo θ_i .(King, 1987)

1.c

θ é o parâmtro que resume certa informação do conjunto de observações da variável dependente. Creio que θ_i, possui o i por poder corresponder a diferentes observações, e o α,não, por se tratar de um parâmetro auxiliar (King, 1987: 10)

1.d

Dizer que Y_i é uma variável aleatória significa que ela foi produzida a partir de dados randomizado, não ao acaso. Mas que foi produzida de forma independente, cada uma de suas observações.

Quando utilizamos o y_i, em minúsculo, estamos nos referindo as observações que compõem nossa variável dependente.

1.e

x_i é um dos possíveis valores que a variável indpendente X pode assumir no modelo estatísitico. E β siginifica o efeito e impacto de X na nossa variável dependente.

2

$$P(trabalhar|C, M) = \pi = logit^{-1}(1.336 - 1.576C - 0.004M)$$

Sendo, $C =$ ter filhos , e $M =$ ser casada com a renda do marido em salários mínimos

2.a

Probalidade de uma mulher entre 21 e 30 anos que não tem filhos e nem é casada trabalahr fora:

```
pr2a <- invlogit(1.336 - 1.576*0 - 0.004 * 0)
pr2a
```

```
## [1] 0.7918314
```

Quando $C = 0$, e $M = 0$, há 79,18% de probabilidade da mulher trabalhar fora.

2.b

Qual a diferença na probabilidade de trabalhar fora quando uma mulher tem filhos?

```
pr2b <- invlogit(1.336 - 1.576*1 - 0.004 * 0) - invlogit(1.336 - 1.576*0 - 0.004 * 0)
pr2b
```

```
## [1] -0.351545
```

A probabilidade de uma mulher solteira e com filhos é 35% menor do que uma mulher solteira e sem filhos.

2.c

Qual a diferença na probabilidade de trabalhar fora entre mães solteiras e mães casadas com maridos que ganham 2 SM?

Ou seja, na primeira situação, $C = 1$ e $M = 0$; e na segunda, $C = 1$, e $M = 2$:

```
pr2c <- invlogit(1.336 - 1.576*1 - 0.004*0) - invlogit(1.336 - 1.576*1 - 0.004*2)
pr2c
```

```
## [1] 0.001970522
```

A probabilidade de uma mãe solteira trabalhar, comparado a uma mãe casa com renda do marido igual a 2M, é de 0.19%.

3

Sabendo que:

$$logit(Pr) = \beta_0 + \beta_i * x$$

E com as informações de que:

$$Pr(EM|R = 0) = 0.27$$

e

$$Pr(EM|R = 12) = 0.88$$

Sendo EM= terminar Ensino Médio, e R = renda dos pais, em receita com base de R$10.000,00. Ao substituirmos as informações na equação de logit, temos:

$$logit(Pr(EM|R)) = \beta_0 + \beta_1 * R$$

Para encontrar β_0, quando $Pr(EM|R = 0) = 0.27$:

$$logit(0.27) = \beta_0 + \beta_1 * 0$$
$$logit(0.27) = \beta_0$$

Logo

```
beta0 <- logit(0.27)
beta0
```

```
## [1] -0.9946226
```

$$\beta_0 = -0.9946$$

Para descobrirmos β_1, quando $Pr(EM|R = 12) = 0.88$ e com β_0 já conhecido:

$$logit(0.88) = -0.9946 + \beta_1 * 12$$

```
beta1 = (logit(0.88) + 0.9946)/12
beta1
```

```
## [1] 0.2489192
```

Logo

$$\beta_1 = 0.2489$$

Por fim, o modelo de regressão logística consiste em:

$$logit(Pr(EM|R)) = -0.9946 + 0.2489 * R$$

4

Seguindo a distribuição da probabilidade de Bernoulli:

$$p(y_i) = Pr(Y_i = y_i) = \pi^{y_i}(1-\pi)^{1-y_i}$$

Para $y_i = 0$:

$$p(0) = Pr(Y_i = 0) = \pi^0(1-\pi)^{1-0}p(0) = 1(1-\pi)p(0) = 1-\pi$$

Para $y_i = 1$:

$$p(1) = Pr(Y_i = 1) = \pi^1(1-\pi)^{1-1}p(1) = \pi(1-\pi)^0p(1) = \pi$$

5

Quando a $\pi_i = 0.5$, temos a região da curva de probabilidade onde esta se encontra mais linear. É o ponto de inclinação máxima, onde a relação entre X e πé quase linear. Logo temos que :

$$logit(0.5) = \alpha + \beta x 0 = \alpha + \beta x$$

Ao derivarmos a reta tangencial no ponto em que $\pi = 0.5$, e substituirmos o resultado na equação anterior, temos que:

$$\beta[\epsilon^{\alpha+\beta x_i}/(1+\epsilon^{\alpha+\beta x_i})^2]\beta[\epsilon^0/(1+\epsilon^0)^2]\beta[1/2^2]\beta/4$$

6

```
folha <- fread("dataFolha.csv", encoding = "UTF-8")
head(folha)
```

```
##       V1 bozo           idade      sexo         partido
## 1:  9132    1 60 anos ou mais Masculino             PSL
## 2:  7765    0 60 anos ou mais Masculino             PDT
## 3:  8673    0    35 a 44 anos Masculino              PT
## 4: 10362    1 60 anos ou mais Masculino             MDB
## 5: 10295    1    25 a 34 anos  Feminino Nenhum/ não tem
## 6:  6173    0    45 a 59 anos  Feminino Nenhum/ não tem
##                                                religiao
## 1:                                             Católica
## 2:                    Não tem religião nenhuma / Agnóstico
## 3: Umbanda, Candomblé ou outras religiões afro-brasileiras
## 4:                                             Católica
## 5:                     Espírita Kardecista, espiritualista
## 6:                                 Evangélica Pentecostal
##                                  escola        rendaf   regiao  raca
## 1:                     Colegial completo    Até 2 S.M.  Sudeste Negra
## 2:        Analfabeto/ primario incompleto    Até 2 S.M.  Sudeste Negra
```

```
## 3: Primario completo/ Ginasial incompleto De 3 a 5 S.M.  Sudeste Negra
## 4:                      Colegial completo De 3 a 5 S.M.  Sudeste Negra
## 5:                      Ginasial completo    Até 2 S.M. Nordeste Negra
## 6:                      Ginasial completo De 3 a 5 S.M.  Sudeste Negra
```

```r
summary(folha)
```

```
##        V1             bozo          idade               sexo
##  Min.   :    5   Min.   :0.000   Length:1000        Length:1000
##  1st Qu.: 2668   1st Qu.:0.000   Class :character   Class :character
##  Median : 5563   Median :0.000   Mode  :character   Mode  :character
##  Mean   : 5511   Mean   :0.365
##  3rd Qu.: 8166   3rd Qu.:1.000
##  Max.   :10926   Max.   :1.000
##    partido            religiao            escola             rendaf
##  Length:1000        Length:1000        Length:1000        Length:1000
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    regiao              raca
##  Length:1000        Length:1000
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

```r
folha$bozo <- as.factor(folha$bozo)
folha$idade <- as.factor(folha$idade)
folha$sexo <- as.factor(folha$sexo)
folha$partido <- as.factor(folha$partido)
folha$regiao <- as.factor(folha$regiao)
folha$escola <- as.factor(folha$escola)
folha$rendaf <- as.factor(folha$rendaf)
folha$raca <- as.factor(folha$raca)
folha$religiao <- as.factor(folha$religiao)

folha <- folha %>%
  mutate(religiaoUnido = case_when( religiao == "  Evangélica Pentecostal" ~ "evangelica",
                                    religiao == "  Evangélica Tradicional" ~ "evangelica",
                                    religiao == "  Evangélica Neo Pentecostal" ~ "evangelica",
                                    religiao == "  Outras Evangélicas" ~ "evangelica",
                                    religiao == "Não tem religião nenhuma / Agnóstico" ~ "ateu",
                                    religiao == "É ateu/ não acredita em Deus" ~ "ateu",
                                    religiao == "Católica" ~ "catolica",
                                    religiao == "Espírita Kardecista, espiritualista" ~ "espiritismo_ou
                                    religiao == "Umbanda, Candomblé ou outras religiões afro-brasileiras
                                    religiao == "Outra religião" ~ "Outra religiao",
                                    religiao == "Judaica" ~ "Outra religiao"))
folha$religiaoUnido <- as.factor(folha$religiaoUnido)

xtabs(~ bozo, data = folha) #menos da metade votou bozo
```

```
## bozo
##   0   1
## 635 365
```

```
xtabs(~ bozo + religiaoUnido, data = folha)
```

```
##     religiaoUnido
## bozo ateu catolica espiritismo_ou_matriz_africana evangelica Outra religiao
##    0   62      319                              48        190             16
##    1   25      171                              27        133              9
```

```
xtabs(~bozo + sexo, data = folha)
```

```
##     sexo
## bozo Feminino Masculino
##    0      395       240
##    1      154       211
```

```
xtabs(~ bozo + raca, data = folha)
```

```
##     raca
## bozo Amarela Branca Indigena Não sabe Negra Outras
##    0      18    206        9        3   388     11
##    1      16    140        4        2   201      2
```

```
xtabs(~ bozo + escola, data = folha)
```

```
##     escola
## bozo Analfabeto/ primario incompleto Colegial completo Colegial incompleto
##    0                             74               193                  62
##    1                             16               138                  37
##     escola
## bozo Ginasial completo Pós graduação Primario completo/ Ginasial incompleto
##    0               56            25                                    102
##    1               26            23                                     30
##     escola
## bozo Superior completo Superior incompleto
##    0               75                  48
##    1               57                  38
```

```
xtabs(~ bozo + idade, data = folha)
```

```
##     idade
## bozo 16 a 24 anos 25 a 34 anos 35 a 44 anos 45 a 59 anos 60 anos ou mais
##    0          102          125          127          161             120
##    1           51           80           75           89              70
```

```r
xtabs(~ bozo + rendaf, data = folha)
```

```
##     rendaf
## bozo Até 2 S.M. De 10 a 20 S.M. De 2 a 3 S.M. De 20 a 50 S.M. De 3 a 5 S.M.
##    0        280              27           110               6            97
##    1         99              30            82               9            78
##     rendaf
## bozo De 5 5 a 10 S.M. Mais de 50 S.M. Não sabe Recusa
##    0               76               0       33      6
##    1               59               2        2      4
```

```r
xtabs(~ bozo + regiao, data = folha)
```

```
##     regiao
## bozo Centro Oeste Nordeste Norte Sudeste Sul
##    0          101      154    27     327  26
##    1           88       44    14     198  21
```

Unimos as religões sobre as principais denominações para diminuir os níveis desta variável. Observando as frequências, podemos ter um pouco ideia sobre a base e como cada grupo se comportou com relação ao voto em Bolsonaro, e termos algum indício de possíveis relações entre as variáveis. Vimos que mais da metade da amostra não votou no atual presidente. Os mais favoráveis são do sexo masculino. Das outras variáveis, aparentemente, a tendência é o voto contra-bolsonaro. Analisaremos melhor isto conforme a análise dos modelos de regressão logística.

*Testando sem interações

```r
fit1_6 <- glm(bozo ~ sexo, data = folha, family = "binomial")
fit2_6 <- glm(bozo ~ sexo + religiaoUnido, data = folha, family = "binomial")
fit3_6 <- glm(bozo ~ sexo + religiaoUnido + raca , data = folha, family = "binomial")
fit4_6 <- glm(bozo ~ sexo + religiaoUnido + raca  + regiao, data = folha, family = "binomial")
fit5_6 <- glm(bozo ~ sexo + religiaoUnido + raca  + regiao + escola, data = folha, family = "binomial")
fit6_6 <- glm(bozo ~ sexo + religiaoUnido + raca  + regiao + escola + rendaf, data = folha, family = "b
fit7_6 <- glm(bozo ~ sexo + religiaoUnido + raca  + regiao + escola + rendaf + idade, data = folha, fami

stargazer(fit1_6, fit2_6, fit3_6, fit4_6, fit5_6, fit6_6, fit7_6, type = "text",

          no.space = TRUE, # to remove the spaces after each line of coefficients

          column.sep.width = "3pt", # to reduce column width

          font.size = "small" # to make font size smaller
          )
```

```
##
## ==========================================================================================
##                                                  Dependent variable:
##                                     ---------------------------------------------
##                                                        bozo
##                                        (1)       (2)       (3)       (4)       (5)       (6)
## ------------------------------------------------------------------------------------------
## sexoMasculino                        0.813***  0.842***  0.859***  0.883***  0.914***  0.840*
```

```
##                                                 (0.134)   (0.135)   (0.136)   (0.139)   (0.142)   (0.14
## religiaoUnidocatolica                                      0.357     0.369     0.492*    0.688**  0.741*
##                                                           (0.260)   (0.262)   (0.266)   (0.274)   (0.28
## religiaoUnidoespiritismo_ou_matriz_africana                0.504     0.485     0.495     0.507     0.50
##                                                           (0.345)   (0.347)   (0.350)   (0.356)   (0.36
## religiaoUnidoevangelica                                   0.656**   0.692**   0.799***  1.011***  1.154*
##                                                           (0.268)   (0.271)   (0.275)   (0.285)   (0.29
## religiaoUnidoOutra religiao                                0.285     0.359     0.403     0.441     0.50
##                                                           (0.488)   (0.493)   (0.499)   (0.510)   (0.51
## racaBranca                                                          -0.200    -0.200    -0.179    -0.18
##                                                                     (0.369)   (0.375)   (0.381)   (0.38
## racaIndigena                                                        -0.810    -0.881    -0.693    -0.67
##                                                                     (0.705)   (0.713)   (0.724)   (0.73
## racaNão sabe                                                        -0.280    -0.255    -0.082    -0.04
##                                                                     (0.990)   (1.003)   (1.039)   (1.01
## racaNegra                                                           -0.524    -0.504    -0.433    -0.41
##                                                                     (0.362)   (0.367)   (0.373)   (0.37
## racaOutras                                                         -1.703**  -1.582*   -1.283    -1.31
##                                                                     (0.854)   (0.860)   (0.868)   (0.88
## regiaoNordeste                                                                -1.178*** -1.102*** -0.965
##                                                                               (0.232)   (0.237)   (0.24
## regiaoNorte                                                                   -0.668*   -0.666*   -0.47
##                                                                               (0.375)   (0.386)   (0.39
## regiaoSudeste                                                                 -0.371**  -0.339*   -0.31
##                                                                               (0.178)   (0.183)   (0.18
## regiaoSul                                                                     -0.196    -0.114     0.01
##                                                                               (0.343)   (0.352)   (0.36
## escolaColegial completo                                                                 1.242***  1.060*
##                                                                                         (0.310)   (0.31
## escolaColegial incompleto                                                               1.103***  0.918
##                                                                                         (0.360)   (0.36
## escolaGinasial completo                                                                  0.811**   0.74
##                                                                                         (0.380)   (0.38
## escolaPós graduação                                                                     1.591***  1.01
##                                                                                         (0.423)   (0.40
## escolaPrimario completo/ Ginasial incompleto                                             0.340     0.29
##                                                                                         (0.358)   (0.36
## escolaSuperior completo                                                                 1.270***  0.818
##                                                                                         (0.344)   (0.37
## escolaSuperior incompleto                                                               1.417***  1.145*
##                                                                                         (0.370)   (0.38
## rendafDe 10 a 20 S.M.                                                                              0.79
##                                                                                                    (0.34
## rendafDe 2 a 3 S.M.                                                                                0.45
##                                                                                                    (0.20
## rendafDe 20 a 50 S.M.                                                                              1.29
##                                                                                                    (0.60
## rendafDe 3 a 5 S.M.                                                                                0.53
##                                                                                                    (0.2
## rendafDe 5 5 a 10 S.M.                                                                             0.52
##                                                                                                    (0.24
## rendafMais de 50 S.M.                                                                              14.04
##                                                                                                   (377.3
## rendafNão sabe                                                                                    -1.913
```

```
##                                                                                                  (0.75
## rendafRecusa                                                                                      0.25
##                                                                                                  (0.7
## idade25 a 34 anos
##
## idade35 a 44 anos
##
## idade45 a 59 anos
##
## idade60 anos ou mais
##
## Constant                                        -0.942*** -1.391*** -1.012**   -0.690   -1.990*** -2.19
##                                                  (0.095)   (0.255)   (0.428)   (0.450)   (0.550)   (0.5
## ------------------------------------------------------------------------------------------------------
## Observations                                       1,000     1,000     1,000     1,000     1,000     1,0
## Log Likelihood                                  -637.472  -633.546  -628.329  -613.551  -594.763  -580.
## Akaike Inf. Crit.                               1,278.944 1,279.092 1,278.658 1,257.103 1,233.525 1,220
## ======================================================================================================
## Note:                                                                                  *p<0.1; **p<0
```

Pertencer ao sexo masculino se mostrou significante e positivo em todos os modelos. Tal qual a religão Evangélica, e com um efeito alto. Dentre as raças, nenhuma se mostrou significativa. Em região, nordeste apresenta um efeito alto com sinal negativo, o que significa menor probabilidade da pessoa votar no Bolsonaro, sendo nordestina, comparado as demais localidades. Em escolaridade, todas se mostraram significantes e positivas, e com uma magnitude alta, de $\beta > 1$. A renda se mostrou significativa acima de 5 S.M.

- Testando novo modelo e com interações

```
fit8_6 <- glm(bozo ~ regiao + sexo + escola + sexo*escola, data = folha, family = "binomial")
display(fit8_6)
```

```
## glm(formula = bozo ~ regiao + sexo + escola + sexo * escola,
##     family = "binomial", data = folha)
##                                                            coef.est coef.se
## (Intercept)                                                -2.27     0.61
## regiaoNordeste                                             -1.07     0.24
## regiaoNorte                                                -0.60     0.38
## regiaoSudeste                                              -0.34     0.18
## regiaoSul                                                  -0.01     0.35
## sexoMasculino                                               1.86     0.69
## escolaColegial completo                                     1.99     0.62
## escolaColegial incompleto                                   1.57     0.68
## escolaGinasial completo                                     1.22     0.72
## escolaPós graduação                                         2.13     0.71
## escolaPrimario completo/ Ginasial incompleto                1.17     0.67
## escolaSuperior completo                                     2.06     0.65
## escolaSuperior incompleto                                   2.11     0.67
## sexoMasculino:escolaColegial completo                      -1.13     0.73
## sexoMasculino:escolaColegial incompleto                    -0.56     0.82
## sexoMasculino:escolaGinasial completo                      -0.49     0.86
## sexoMasculino:escolaPós graduação                          -0.91     0.92
## sexoMasculino:escolaPrimario completo/ Ginasial incompleto -1.22     0.81
## sexoMasculino:escolaSuperior completo                      -1.35     0.78
```

```
## sexoMasculino:escolaSuperior incompleto                    -1.19     0.82
## ---
##   n = 1000, k = 20
##   residual deviance = 1203.9, null deviance = 1312.5 (difference = 108.6)
```

```
fit9_6 <- glm(bozo ~ sexo + escola + sexo*escola, data= folha, family =  "binomial")
display(fit9_6)
```

```
## glm(formula = bozo ~ sexo + escola + sexo * escola, family = "binomial",
##     data = folha)
##                                                           coef.est coef.se
## (Intercept)                                               -2.66     0.60
## sexoMasculino                                              1.79     0.68
## escolaColegial completo                                    1.94     0.62
## escolaColegial incompleto                                  1.54     0.68
## escolaGinasial completo                                    1.16     0.71
## escolaPós graduação                                        2.17     0.71
## escolaPrimario completo/ Ginasial incompleto               1.21     0.66
## escolaSuperior completo                                    2.19     0.64
## escolaSuperior incompleto                                  2.14     0.66
## sexoMasculino:escolaColegial completo                     -1.01     0.72
## sexoMasculino:escolaColegial incompleto                   -0.58     0.81
## sexoMasculino:escolaGinasial completo                     -0.39     0.85
## sexoMasculino:escolaPós graduação                         -0.76     0.92
## sexoMasculino:escolaPrimario completo/ Ginasial incompleto -1.27     0.80
## sexoMasculino:escolaSuperior completo                     -1.35     0.77
## sexoMasculino:escolaSuperior incompleto                   -1.10     0.82
## ---
##   n = 1000, k = 16
##   residual deviance = 1228.3, null deviance = 1312.5 (difference = 84.2)
```

```
fit10_6 <- glm(bozo ~ sexo + raca+ escola + sexo*raca,  data= folha, family =  "binomial")
display(fit10_6)
```

```
## glm(formula = bozo ~ sexo + raca + escola + sexo * raca, family = "binomial",
##     data = folha)
##                                              coef.est coef.se
## (Intercept)                                  -1.23     0.56
## sexoMasculino                                 0.08     0.71
## racaBranca                                   -0.58     0.51
## racaIndigena                                 -1.27     1.22
## racaNão sabe                                 15.02   604.86
## racaNegra                                    -0.92     0.51
## racaOutras                                   -0.96     1.27
## escolaColegial completo                       1.24     0.30
## escolaColegial incompleto                     1.08     0.35
## escolaGinasial completo                       0.80     0.37
## escolaPós graduação                           1.48     0.41
## escolaPrimario completo/ Ginasial incompleto  0.36     0.35
## escolaSuperior completo                       1.28     0.34
## escolaSuperior incompleto                     1.38     0.36
## sexoMasculino:racaBranca                      0.71     0.74
## sexoMasculino:racaIndigena                    1.23     1.54
```

```
## sexoMasculino:racaNão sabe                    -29.42   782.63
## sexoMasculino:racaNegra                         0.94     0.73
## sexoMasculino:racaOutras                       -0.72     1.74
## ---
##   n = 1000, k = 19
##   residual deviance = 1215.5, null deviance = 1312.5 (difference = 97.0)
```

Observamos que as variáveis sozinhas tem significância, mas a perdem quando em interação. Até o sinal se mostra diferente do que esperaríamos, o impacto de ser dos sexo masculino e variações na escolaridade se mostrou negativo.

Analisando o residual deviance:

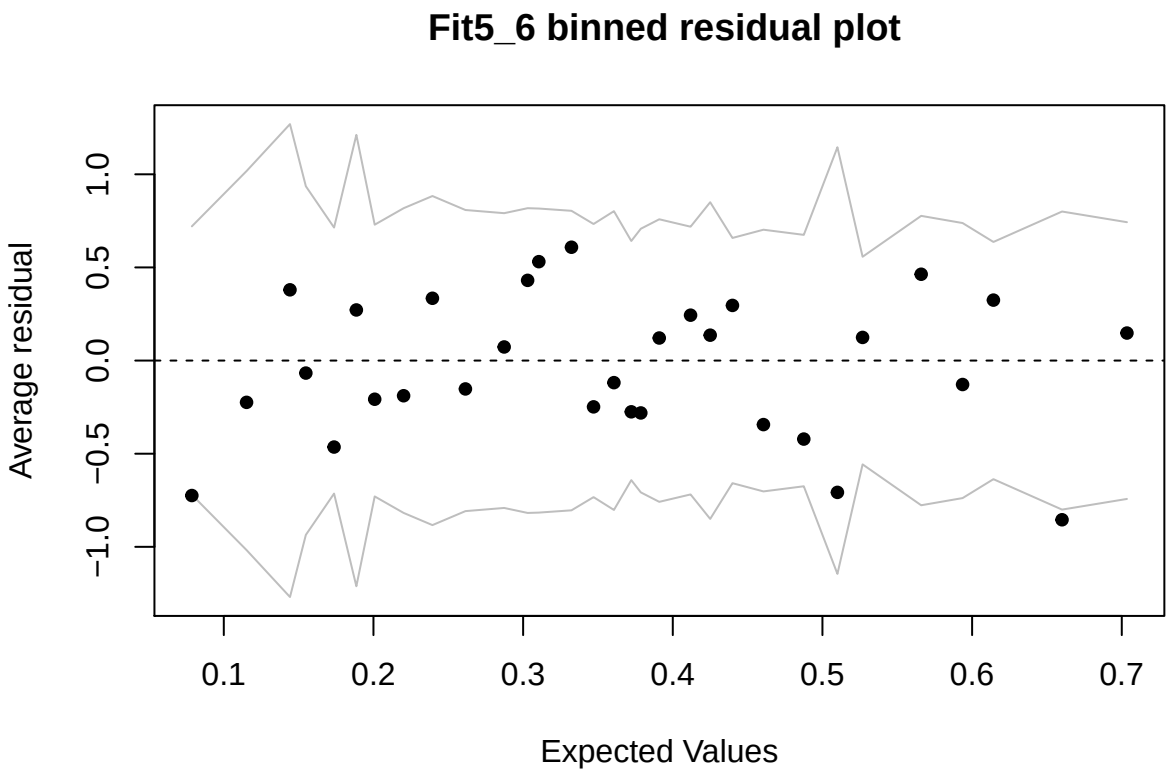
```
tibble::tibble( Modelo = c("Intercepto","Fit1_6", "Fit2_6", "Fit3_6","Fit4_6", 'fit5_6', 'fit6_6',' fit
                Deviance = c(fit1_6$null.deviance,fit1_6$deviance, fit2_6$deviance,
                             fit3_6$deviance, fit4_6$deviance,
                             fit5_6$deviance, fit6_6$deviance,
                             fit7_6$deviance, fit8_6$deviance,
                             fit9_6$deviance, fit10_6$deviance)) %>%
  qflextable()
```

```
## Warning: Warning: fonts used in ‘flextable‘ are ignored because the ‘pdflatex‘
## engine is used and not ‘xelatex‘ or ‘lualatex‘. You can avoid this warning
## by using the ‘set_flextable_defaults(fonts_ignore=TRUE)‘ command or use a
## compatible engine by defining ‘latex_engine: xelatex‘ in the YAML header of the
## R Markdown document.
```

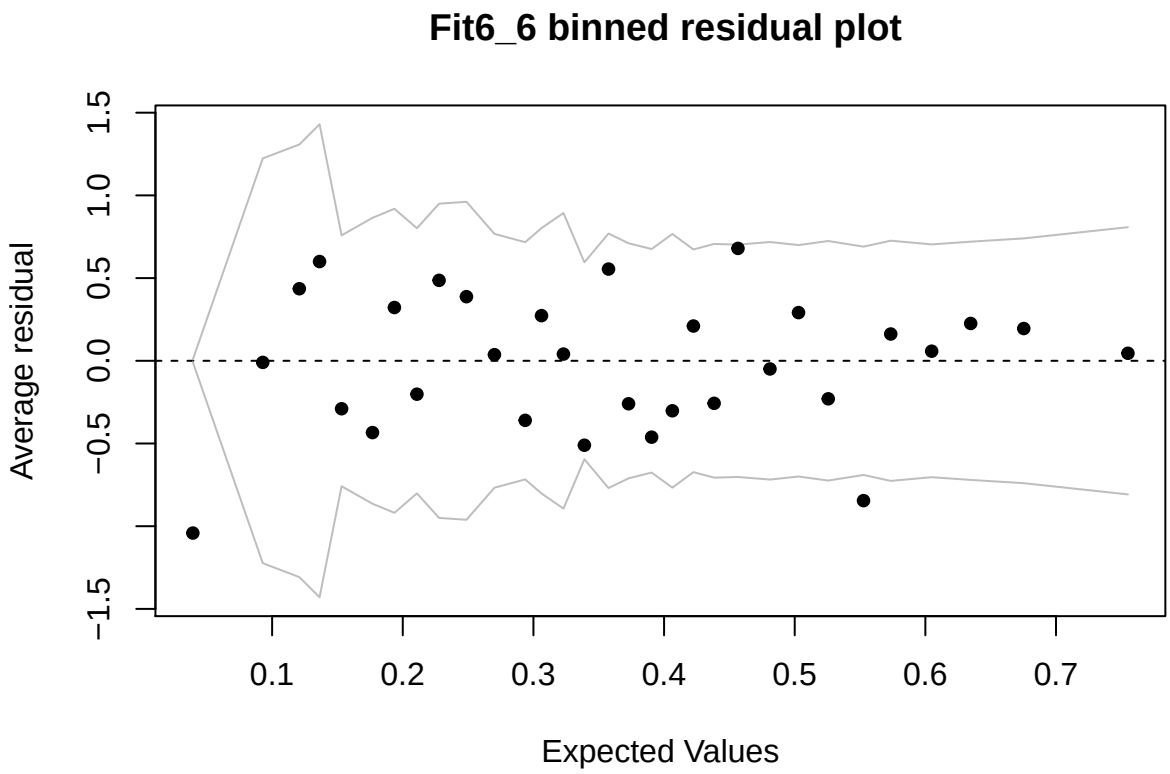
Modelo	Deviance
Intercepto	1,312.482
Fit1_6	1,274.944
Fit2_6	1,267.092
Fit3_6	1,256.658
Fit4_6	1,227.103
fit5_6	1,189.525
fit6_6	1,160.371
fit7_6	1,155.542
fit8_6	1,203.877
fit9_6	1,228.252
fit10_6	1,215.465

De todos os modelos rodados, percebemos uma queda maior no residual deviance de fit5_6, fit_6, fit7_6. Este último inclui todas as variáveis, logo, é bem possível que tal redução e seja atribuido a isso. O fit10_6 possui menos variáveis, e tem uma queda considerável na residual deviance. O problema deste é que possui 2 erros padrões com um valor muito alto. Nesse sentido, vamos comparar os gráficos de resíduos de fit5_6, fit_6, fit10_6 e decidirmos pelo melhor.

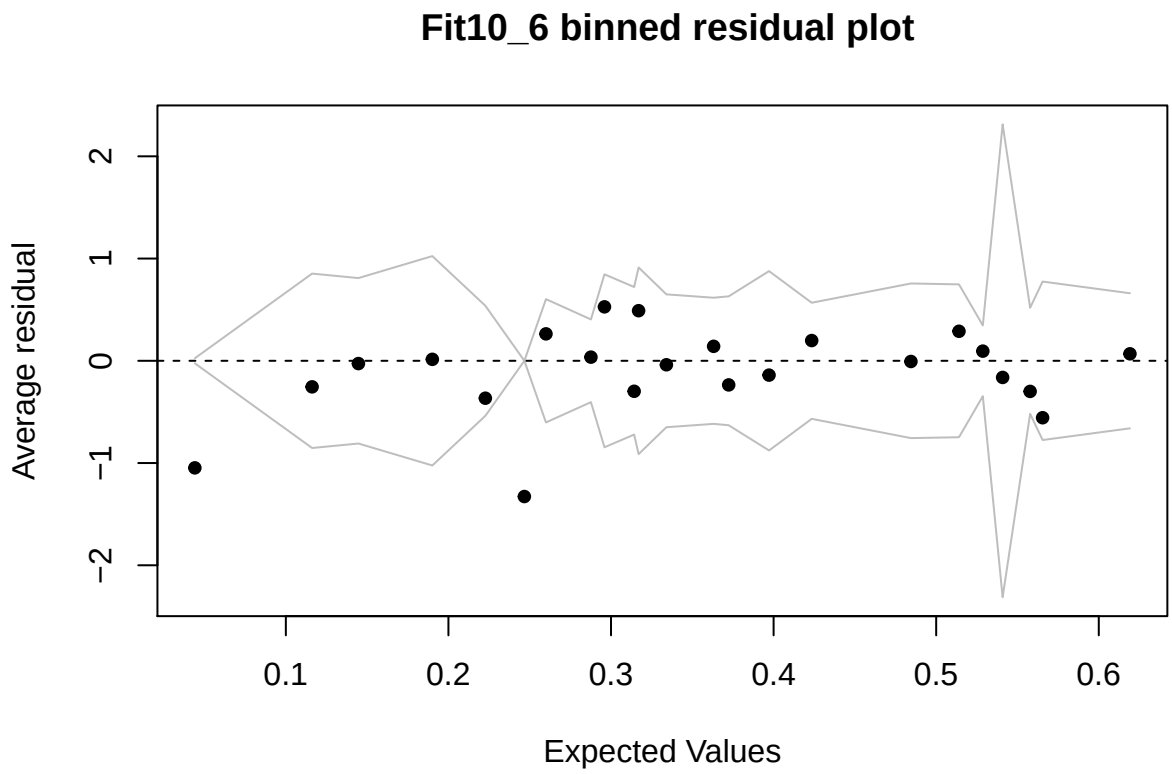
```
binnedplot(x = fit5_6$fitted.values,y = fit5_6$residuals,  main=" Fit5_6 binned residual plot")
```




```
binnedplot(x = fit6_6$fitted.values,y = fit6_6$residuals, main=" Fit6_6 binned residual plot")
```




```
binnedplot(x = fit10_6$fitted.values,y = fit10_6$residuals, main=" Fit10_6 binned residual plot")
```



Analisando os resíduos, os que se encontram mais perto da média zero são os encontrados no modelo fit5_6.

```
invlogit(coefficients(fit5_6))
```

```
##                                   (Intercept)
##                                     0.1202929
##                                   sexoMasculino
##                                     0.7138699
##                           religiaoUnidocatolica
##                                     0.6655636
##  religiaoUnidoespiritismo_ou_matriz_africana
##                                     0.6240852
##                          religiaoUnidoevangelica
##                                     0.7332751
##                    religiaoUnidoOutra religiao
##                                     0.6084737
##                                    racaBranca
##                                     0.4554366
##                                   racaIndigena
##                                     0.3333477
##                                   racaNão sabe
##                                     0.4796071
##                                     racaNegra
##                                     0.3933608
##                                    racaOutras
##                                     0.2170396
```

```
##                                regiaoNordeste
##                                     0.2493772
##                                   regiaoNorte
##                                     0.3394766
##                                 regiaoSudeste
##                                     0.4160055
##                                     regiaoSul
##                                     0.4715167
##                       escolaColegial completo
##                                     0.7758957
##                     escolaColegial incompleto
##                                     0.7508771
##                       escolaGinasial completo
##                                     0.6924102
##                           escolaPós graduação
##                                     0.8307505
## escolaPrimario completo/ Ginasial incompleto
##                                     0.5841653
##                       escolaSuperior completo
##                                     0.7806837
##                     escolaSuperior incompleto
##                                     0.8049381
```

A probabilidade de votar no Bolsonaro aumenta em 70% se a pessoa for homem, quando comparado a ser mulher. Há o aumento de 73% nesse voto se sua religiosidade for evangélica, comparada as demais. Se for branco, há 45% de probabilidade a mais no voto, em relações as outras. Já ser da região Nordeste apresenta um impacto negativo, de menos 24% de votar em Bolsonaro, comparado as outras regiões. Já possuir o nível educacional de ensino superior completo aumenta 78% a probabilidade do voto em Bolsonaro, comparado as outras.

7

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.0.5
```

```
## Registered S3 methods overwritten by 'car':
##   method                          from
##   influence.merMod                lme4
##   cooks.distance.influence.merMod lme4
##   dfbeta.influence.merMod         lme4
##   dfbetas.influence.merMod        lme4
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some

## The following object is masked from 'package:arm':
##
##     logit
```

```
data("Womenlf")

# transformar a variavel trabalho em binária

trabMulher <- Womenlf %>%
  mutate(trabalho = case_when( partic == "fulltime" ~ 1,
                               partic == "parttime" ~ 1,
                               partic == "not.work" ~ 0))
trabMulher <- trabMulher %>%
  mutate(crianca = case_when(children == "absent" ~ 0,
                             children == "present" ~ 1))

#transformando a nova variavel em factor
trabMulher$trabalho <- as.factor(trabMulher$trabalho)
trabMulher$crianca <- as.factor(trabMulher$crianca)

trabMulher$partic <- NULL
trabMulher$children <- NULL
rm(Womenlf)

glimpse(trabMulher)
```

```
## Rows: 263
## Columns: 4
## $ hincome  <int> 15, 13, 45, 23, 19, 7, 15, 7, 15, 23, 23, 13, 9, 9, 45, 15, 5~
## $ region   <fct> Ontario, Ontario, Ontario, Ontario, Ontario, Ontario, Ontario~
## $ trabalho <fct> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1~
## $ crianca  <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0~
```

```
summary(trabMulher)
```

```
##     hincome           region    trabalho crianca
##  Min.   : 1.00   Atlantic: 30   0:155    0: 79
##  1st Qu.:10.00   BC      : 29   1:108    1:184
##  Median :14.00   Ontario :108
##  Mean   :14.76   Prairie : 31
##  3rd Qu.:19.00   Quebec  : 65
##  Max.   :45.00
```

```
#trabalho e criança são variáveis dicotômicas
# renda marital é contínua
# região é discreta

sapply(trabMulher, function(x) sum(is.na(x)))
```

```
##  hincome   region trabalho  crianca
##        0        0        0        0
```

```
# frequencia dado o trabalho
xtabs( ~ trabalho, data = trabMulher)
```

```
## trabalho
##   0   1
## 155 108
```

```
xtabs(~ trabalho + region, data = trabMulher)
```

```
##         region
## trabalho Atlantic BC Ontario Prairie Quebec
##        0       20 14      64      17     40
##        1       10 15      44      14     25
```

```
xtabs(~ trabalho + crianca, data = trabMulher)
```

```
##         crianca
## trabalho   0   1
##        0  26 129
##        1  53  55
```

```
#talvez seja melhor dividir a renda do marido
trabMulher$hincome5 <- trabMulher$hincome/5

glimpse(trabMulher)
```

```
## Rows: 263
## Columns: 5
## $ hincome  <int> 15, 13, 45, 23, 19, 7, 15, 7, 15, 23, 23, 13, 9, 9, 45, 15, 5~
## $ region   <fct> Ontario, Ontario, Ontario, Ontario, Ontario, Ontario, Ontario~
## $ trabalho <fct> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1~
## $ crianca  <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0~
## $ hincome5 <dbl> 3.0, 2.6, 9.0, 4.6, 3.8, 1.4, 3.0, 1.4, 3.0, 4.6, 4.6, 2.6, 1~
```

- Rodando as regressões logísticas

```
fit1_7 <- glm(trabalho ~ crianca, data = trabMulher, family = "binomial")
fit2_7 <- glm(trabalho ~ crianca + hincome, data = trabMulher, family = "binomial")
fit3_7 <- glm(trabalho ~ crianca + hincome + region, data = trabMulher, family = "binomial")
fit4_7 <- glm( trabalho ~ hincome + crianca + hincome*crianca, data = trabMulher, family = "binomial" )
fit5_lm_7 <- lm(trabalho ~crianca + hincome + crianca:hincome, data = trabMulher, family = "binomial" )
```

```
## Warning in model.response(mf, "numeric"): using type = "numeric" with a factor
## response will be ignored
```

```
## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :
##  extra argument 'family' will be disregarded
```

```
## Warning in Ops.factor(y, z$residuals): '-' not meaningful for factors
```

```
stargazer(fit1_7, fit2_7, fit3_7, fit4_7, fit5_lm_7,
          type = "text")
```

```
##
## ===============================================================
##                                 Dependent variable:
##                     ------------------------------------------
##                                       trabalho
##                                    logistic                   OLS
##                        (1)       (2)       (3)       (4)     (5)
## ---------------------------------------------------------------
## crianca1            -1.565*** -1.576*** -1.604*** -2.046*** -0.481
##                      (0.289)   (0.292)   (0.302)   (0.677)
##
## hincome:crianca1                                    0.032
##                                                    (0.041)
##
## hincome                        -0.042**  -0.045**   -0.062*  -0.014
##                                 (0.020)   (0.021)   (0.033)
##
## regionBC                                   0.342
##                                           (0.585)
##
## regionOntario                              0.188
##                                           (0.468)
##
## regionPrairie                              0.472
##                                           (0.557)
##
## regionQuebec                              -0.173
##                                           (0.500)
##
## crianca1:hincome                                              0.008
##
##
## Constant            0.712***  1.336***   1.268**  1.640***  1.870
##                      (0.239)   (0.384)   (0.553)   (0.558)
##
## ---------------------------------------------------------------
## Observations           263       263       263       263     263
## Log Likelihood      -162.279  -159.866  -158.651  -159.562
## Akaike Inf. Crit.    328.559   325.733   331.301   327.124
## ===============================================================
## Note:                                *p<0.1; **p<0.05; ***p<0.01
```

No priemiro modelo, fit1_7, tanto o intercepto (mulher trabalhar sem filhos) quanto o coeficiente de ter filhos é estatísticamente significativo. Isto é, coeficiente da estimativa não está na zona de rejeição de 2x a mais ou a menos do coeficiente de erro padrão. O coeficente sobre ter filho apresenta um sinal negativo, ou seja, a presença de criança impacta negativamente a mulher trabalhar.

```
invlogit(coef(fit1_7))
```

```
## (Intercept)    crianca1
```

```
##   0.6708861   0.1729769
```

```
coef(fit1_7)[2]/4 # Beta/4 probabilidade de uma mulher trabalahr tendo um filho cai em 39%
```

```
##   crianca1
## -0.3911686
```

Utilizando a função invlogit(), conseguimos melhor interpretar a regressão logística. Neste caso, a probabilidade de uma mulher sem filhos trabalhar é de 67%. Utilizando o $\beta/4$, podemos, também, concluir que a probabilidade de uma mulher trabalhar, tendo filhos, cai 39,11%.

No Fit2_7, a renda marital foi acrescentada ao modelo. O sinal desta variável indica que ela impacta negativamente a mulher trabalhar. Espera-se que, defato, o sinal apontasse esse sentido. Em um casal com maior renda, vindo do marido, as probabilidade da mulher estar no mercado de trabalho tende, em tese, a ser menor. Porém, esperávamos que o efeito apresentasse uma magnitude maior.

```
invlogit(coef(fit2_7))
```

```
## (Intercept)    crianca1     hincome
##   0.7918033   0.1714127   0.4894245
```

```
coef(fit2_7)[3]/4
```

```
##     hincome
## -0.01057711
```

Calculando o impacto desta variável, a diferença máxima entre uma mulher solteira e sem filhos dado um acréscimo no salário do marido é de uma queda de 10% na probabilidade dela trabalhar.

Em Fit3_7, foi adicionado a variável região. Como é perceptível, os coeficientes de erro padrão de cada uma das categorias desta variável é muito grande, logo, não estatisticamente significante. Intercepto e ter filho permanecem significativos.

Em fit4_7, testando a interação entre ter filho e o acréscimo na renda do marido. A renda do marido permanece com pouca significância. A interação entre as duas variáveis também é de baixa magnitude e sem significância estatística.

Por último, no modelo fit5_lm_7, onde aplicamos um modelo linear, a interação entre criança e renda do marido não obteve significância.

Os resultodos do efeito de ter criança e renda marital se mostram um tanto contra intuitivos, sobre o trabalho feminino. Podemos considerar que o tamanho reduzido da amostra esteja propriciando isto.

```
tibble::tibble( Modelo = c("Intercepto","Fit1_7", "Fit2_7", "Fit3_7","Fit4_7"),
                Deviance = c(fit1_7$null.deviance,fit1_7$deviance, fit2_7$deviance,
                             fit3_7$deviance, fit4_7$deviance)) %>%
  qflextable()
```

```
## Warning: Warning: fonts used in ‘flextable‘ are ignored because the ‘pdflatex‘
## engine is used and not ‘xelatex‘ or ‘lualatex‘. You can avoid this warning
## by using the ‘set_flextable_defaults(fonts_ignore=TRUE)‘ command or use a
## compatible engine by defining ‘latex_engine: xelatex‘ in the YAML header of the
## R Markdown document.
```

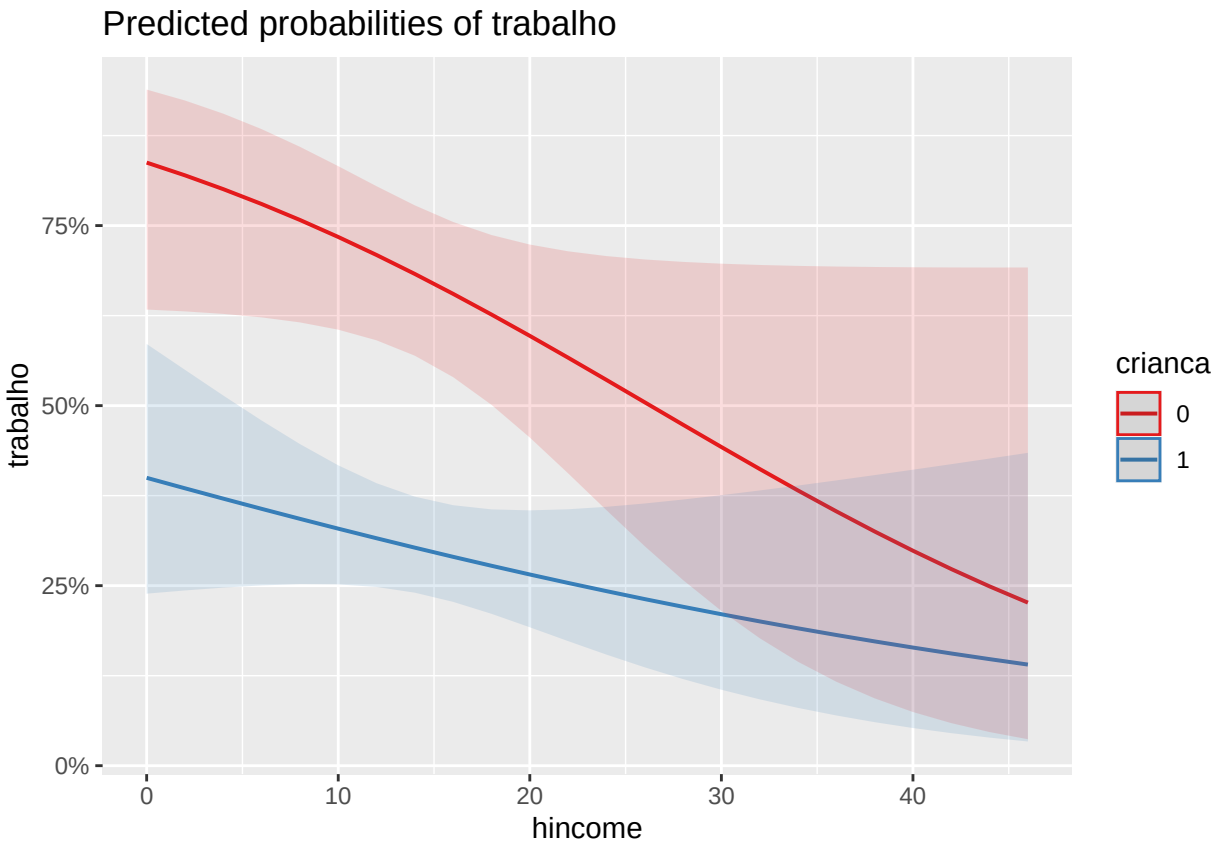
Modelo	Deviance
Intercepto	356.1509
Fit1_7	324.5589
Fit2_7	319.7325
Fit3_7	317.3011
Fit4_7	319.1242

Comparando a residual deviance, o melhor modelo é o segundo modelo (fit2_7), onde as todas as varíaveis e intercepto aparecem significativas.

- Verificando o impacto da renda do Marido para a presença ou não de crianças:

```
sjPlot::plot_model(fit4_7, type = "int")
```

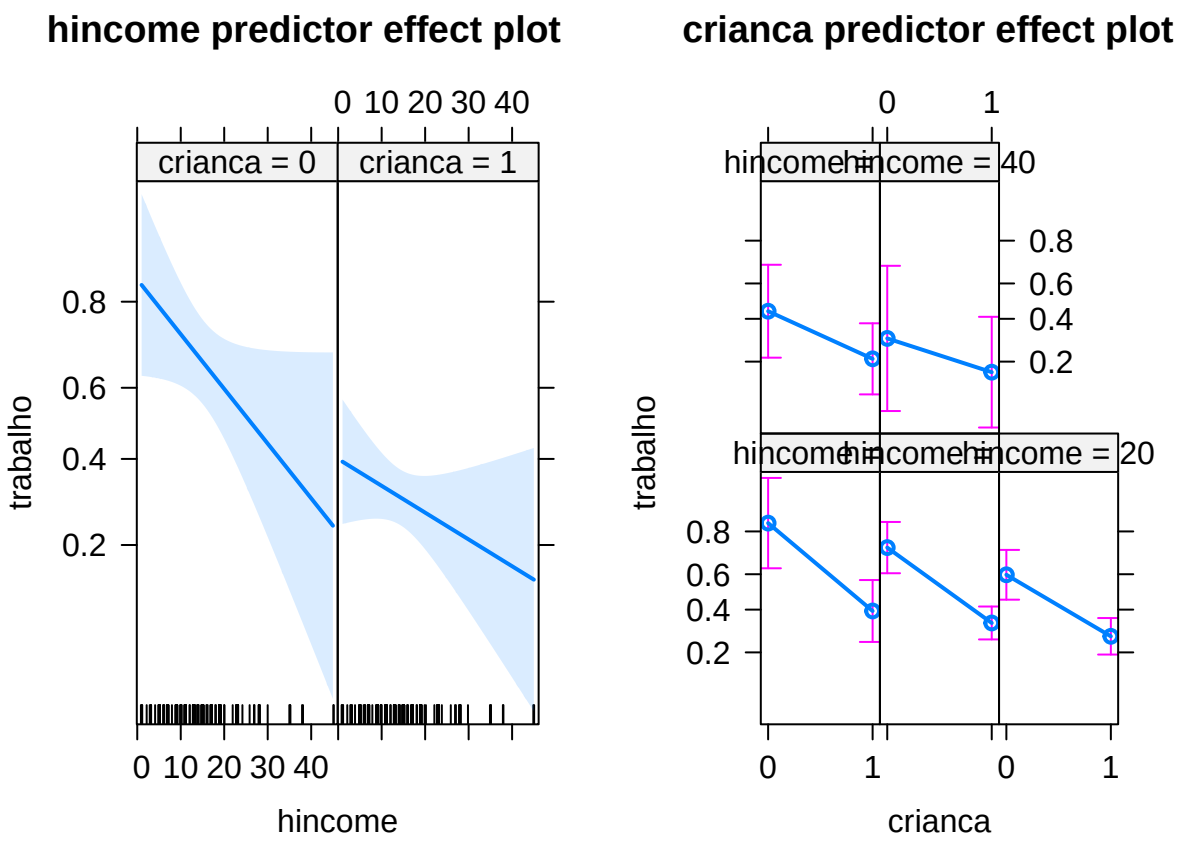
```
## Data were 'prettified'. Consider using 'terms="hincome [all]"' to get smooth plots.
```



Podemos observar no gráfico acima que ter ou não filho diminui as probabilidades de uma mulher trabalhar, dado o aumento na renda do marido. Atribuindo diferentes valores há variável criança, sim ou não, podemos ver que as duas curvas se aproximam conforme cresce a renda marital. Esta afinidade nos dá indicío de haver interação entre marido e criança.

- Usando pacote(Effects)

```
plot(predictorEffects(fit4_7))
```




```
invlogit(0.2)
```

```
## [1] 0.549834
```

Analidando o gráfico acima, consiguimos observar que ter ou não criança, dado o aumento da renda do marido, não afeta, de forma diferente, probabilidade da mulher trabalhar. Um meulher sem filhos, com renda do marido = 40, tem a probabilidade de trabalhar próximo a 54%, tal qual uma mulher com filho e mesma renda marital.
Ou seja, este gráfico reforça a ideia de que não há interação entre essas duas variáveis independentes, como apontado na análise dos coeficientes do modelo.

8

Os dados utilizados são provenientes do pacote(carData) e contém os dados sobre os passageiros do navio Titanic. As informações dizem respeito a status do indivíduo (passengerClass), idade (age), sexo (sex) e sobrevivência ao desastre (survived). Esta última, uma variável dicotômica.

```
data("TitanicSurvival")

dim(TitanicSurvival) #1309 obs
```

```
## [1] 1309    4
```

```
str(TitanicSurvival)
```

```
## 'data.frame':    1309 obs. of  4 variables:
##  $ survived      : Factor w/ 2 levels "no","yes": 2 2 1 1 1 2 2 1 2 1 ...
##  $ sex           : Factor w/ 2 levels "female","male": 1 2 1 2 1 2 1 2 1 2 ...
##  $ age           : num  29 0.917 2 30 25 ...
##  $ passengerClass: Factor w/ 3 levels "1st","2nd","3rd": 1 1 1 1 1 1 1 1 1 1 ...
```

```
summary(TitanicSurvival) #263 NA's em age
```

```
##  survived      sex           age          passengerClass
##  no :809   female:466   Min.   : 0.1667   1st:323
##  yes:500   male  :843   1st Qu.:21.0000   2nd:277
##                         Median :28.0000   3rd:709
##                         Mean   :29.8811
##                         3rd Qu.:39.0000
##                         Max.   :80.0000
##                         NA's   :263
```

```
nrow(TitanicSurvival[is.na(TitanicSurvival$age),])
```

```
## [1] 263
```

```
faltantes<- TitanicSurvival[is.na(TitanicSurvival$age),]
summary(faltantes) #maioria de não sobreviventes
```

```
##  survived      sex           age      passengerClass
##  no :190   female: 78   Min.   : NA   1st: 39
##  yes: 73   male  :185   1st Qu.: NA   2nd: 16
##                         Median : NA   3rd:208
##                         Mean   :NaN
##                         3rd Qu.: NA
##                         Max.   : NA
##                         NA's   :263
```

```
# maioria de homens e de passageiros da terceira classe

titanic.sobrev <- na.omit(TitanicSurvival) # total de 1046 obs restantes
rm(TitanicSurvival)

titanic.sobrev <- titanic.sobrev %>%
  mutate(sobrevivencia = case_when(survived == "yes" ~ 1,
                                   survived == "no" ~ 0),
         sexo = case_when(sex == "female"  ~ 1,
                          sex == "male" ~ 0))
titanic.sobrev <- titanic.sobrev %>%
  mutate(idade = case_when( age <= 20 ~ "jovem",
                            age > 20 & age <= 59 ~ "adulto",
```

```
                            age > 60 ~ "idoso"))

titanic.sobrev$idade <- as.factor(titanic.sobrev$idade)
titanic.sobrev$sobrevivencia <- as.factor(titanic.sobrev$sobrevivencia)
titanic.sobrev$sexo <- as.factor(titanic.sobrev$sexo)
titanic.sobrev <- na.omit(titanic.sobrev)

str(titanic.sobrev)
```

```
## 'data.frame':    1039 obs. of  7 variables:
##  $ survived      : Factor w/ 2 levels "no","yes": 2 2 1 1 1 2 2 1 2 1 ...
##  $ sex           : Factor w/ 2 levels "female","male": 1 2 1 2 1 2 1 2 1 2 ...
##  $ age           : num  29 0.917 2 30 25 ...
##  $ passengerClass: Factor w/ 3 levels "1st","2nd","3rd": 1 1 1 1 1 1 1 1 1 1 ...
##  $ sobrevivencia : Factor w/ 2 levels "0","1": 2 2 1 1 1 2 2 1 2 1 ...
##  $ sexo          : Factor w/ 2 levels "0","1": 2 1 2 1 2 1 2 1 2 1 ...
##  $ idade         : Factor w/ 3 levels "adulto","idoso",..: 1 3 3 1 1 1 2 1 1 2 ...
##  - attr(*, "na.action")= 'omit' Named int [1:7] 41 105 110 268 269 313 414
##   ..- attr(*, "names")= chr [1:7] "Bucknell, Mrs. William Robert (" "Fortune, Mrs. Mark (Mary McDoug
```

```
xtabs(~ sobrevivencia, data = titanic.sobrev) #mais da metade não sobreviveu
```

```
## sobrevivencia
##   0   1
## 616 423
```

```
xtabs( ~ sobrevivencia + sexo, data = titanic.sobrev) #mais homens morreram
```

```
##              sexo
## sobrevivencia   0   1
##             0 521  95
##             1 134 289
```

```
xtabs(~ sobrevivencia + passengerClass, data = titanic.sobrev) #mais sobreviveram na primeira classe, m
```

```
##              passengerClass
## sobrevivencia 1st 2nd 3rd
##             0 102 144 370
##             1 177 115 131
```

```
summary(titanic.sobrev)
```

```
##  survived      sex           age          passengerClass sobrevivencia sexo
##  no :616   female:384   Min.   : 0.1667   1st:279        0:616         0:655
##  yes:423   male  :655   1st Qu.:21.0000   2nd:259        1:423         1:384
##                         Median :28.0000   3rd:501
##                         Mean   :29.6782
##                         3rd Qu.:38.0000
##                         Max.   :80.0000
##     idade
```

```
##  adulto:758
##  idoso : 33
##  jovem :248
##
##
##
```

Em uma primeira análise dos dados, observamos que mais da metade dos passageiros não sobreviveu. Dentre os que conseguiram socorro, mais mulheres e pessoas da primeira classe conseguiram se salvar. Entre os passageiros da terceira classe, mais da metade não sobreviveu.

Agora iremos rodar algumas regressões para tentar explicar a probabilidade de sobrevivência neste contexto:

```
fit1_8 <- glm(sobrevivencia ~ sexo, data = titanic.sobrev, family = "binomial")
display(fit1_8)
```

```
## glm(formula = sobrevivencia ~ sexo, family = "binomial", data = titanic.sobrev)
##             coef.est coef.se
## (Intercept) -1.36     0.10
## sexo1        2.47     0.15
## ---
##   n = 1039, k = 2
##   residual deviance = 1093.4, null deviance = 1404.3 (difference = 310.9)
```

```
fit2_8 <- glm(sobrevivencia ~ sexo + idade, data = titanic.sobrev, family = "binomial")
display(fit2_8)
```

```
## glm(formula = sobrevivencia ~ sexo + idade, family = "binomial",
##     data = titanic.sobrev)
##             coef.est coef.se
## (Intercept) -1.36     0.11
## sexo1        2.46     0.15
## idadeidoso  -0.50     0.47
## idadejovem   0.09     0.18
## ---
##   n = 1039, k = 4
##   residual deviance = 1091.8, null deviance = 1404.3 (difference = 312.5)
```

```
fit3_8<- glm(sobrevivencia ~ sexo + idade + passengerClass, data = titanic.sobrev, family = "binomial")
display(fit3_8)
```

```
## glm(formula = sobrevivencia ~ sexo + idade + passengerClass,
##     family = "binomial", data = titanic.sobrev)
##                   coef.est coef.se
## (Intercept)       -0.32     0.16
## sexo1              2.49     0.17
## idadeidoso        -1.18     0.50
## idadejovem         0.59     0.19
## passengerClass2nd -1.02     0.22
## passengerClass3rd -1.97     0.21
## ---
##   n = 1039, k = 6
##   residual deviance = 990.1, null deviance = 1404.3 (difference = 414.2)
```

```
fit4_8<- glm(sobrevivencia ~ sexo + idade + passengerClass + passengerClass*sexo, data = titanic.sobrev
display(fit4_8)
```

```
## glm(formula = sobrevivencia ~ sexo + idade + passengerClass +
##     passengerClass * sexo, family = "binomial", data = titanic.sobrev)
##                           coef.est coef.se
## (Intercept)               -0.57     0.18
## sexo1                      3.84     0.49
## idadeidoso                -1.26     0.60
## idadejovem                 0.64     0.19
## passengerClass2nd         -1.30     0.29
## passengerClass3rd         -1.22     0.23
## sexo1:passengerClass2nd    0.11     0.64
## sexo1:passengerClass3rd -2.42     0.54
## ---
##   n = 1039, k = 8
##   residual deviance = 941.1, null deviance = 1404.3 (difference = 463.2)
```

O último modelo, fit4_8, embora conte com um pouco de aumento no erro padrão dos coeficientes, aparenta ser o melhor modelo. A única variável sem significância estatística é a interação entre ser mulher e pertencer a segunda classe. Podemos observar também que esse é o modelo com maior queda de deviance, indicando um melhor caráter explicativo as variáveis utilizadas.

```
invlogit(coef(fit4_8))
```

```
##             (Intercept)                    sexo1              idadeidoso
##               0.36092608               0.97888700               0.22110039
##              idadejovem       passengerClass2nd       passengerClass3rd
##               0.65485492               0.21428392               0.22711722
## sexo1:passengerClass2nd sexo1:passengerClass3rd
##               0.52841934               0.08170958
```

Podemos interpretar as probailidades de sobrevivência, $p(y) = 1$, da seguinte forma: a probabilidade de uma mulher sobreviver é de 97%, comparada aos homens. Ser idoso diminui em 13% a chance de sobrevivência, comparada as outras faixa etárias . Sobre o status, um passageiro de terceira classe tem 22% de pribabilidade a menos de sobreviver, comparado as outras classes.

*Calculando a taxa de erro.

```
prev3_8 <- predict(fit3_8, type = "response")
erro3_8 <- mean((prev3_8 > 0.5 & titanic.sobrev$sobrevivencia == 0) | (prev3_8 < .5 & titanic.sobrev$sob
erro3_8
```

```
## [1] 0.2194418
```

```
prev4_8 <- predict(fit4_8, type = "response")
erro4_8 <- mean((prev4_8 > 0.5 & titanic.sobrev$sobrevivencia == 0) | (prev4_8 < .5 & titanic.sobrev$sob
erro4_8
```

```
## [1] 0.2088547
```

O modelo com menor taxa de erro é o fit4_8, com 20,88%. Um erro razoavelmente alto, o que não traz muita confiança sobre a capacidade preditiva do modelo.

```
tibble::tibble( Modelo = c("Intercepto","Fit1_8", "Fit2_8", "Fit3_8","Fit4_8"),
                Deviance = c(fit1_8$null.deviance,fit1_8$deviance, fit2_8$deviance,
                             fit3_8$deviance, fit4_8$deviance)) %>%
  qflextable()
```

```
## Warning: Warning: fonts used in ‘flextable‘ are ignored because the ‘pdflatex‘
## engine is used and not ‘xelatex‘ or ‘lualatex‘. You can avoid this warning
## by using the ‘set_flextable_defaults(fonts_ignore=TRUE)‘ command or use a
## compatible engine by defining ‘latex_engine: xelatex‘ in the YAML header of the
## R Markdown document.
```

Modelo	Deviance
Intercepto	1,404.3000
Fit1_8	1,093.4218
Fit2_8	1,091.8383
Fit3_8	990.1301
Fit4_8	941.1106

Podemos observar que o modelo que apresenta maior queda da deviança é o fit4_8, comparativamente ao modelo nulo (intercepto) e aos demais modelos. Ou seja, é o modelo que melhor explica nossa variável dependente.

- Construção de cenário:

  (1) Probabilidade de sobrevivência quando o passageiro é um homem, jovem e da primeira classe.

```
titanic.sobrev$idade <- relevel(titanic.sobrev$idade, ref = "jovem")
titanic.sobrev$passengerClass <- relevel(titanic.sobrev$passengerClass, ref = "1st")
attach(titanic.sobrev)
coefficients(fit4_8)
```

```
##             (Intercept)                   sexo1              idadeidoso
##               -0.5713470                3.8365271               -1.2592653
##              idadejovem       passengerClass2nd       passengerClass3rd
##               0.6404489               -1.2992936               -1.2246611
## sexo1:passengerClass2nd sexo1:passengerClass3rd
##               0.1138000               -2.4193425
```

```
beta <- coef(fit4_8)

s1 <- 0 # sexo = masculino
i1 <- 1 #idade jovem
ps1 <- 1 # 1º classe

prob <- invlogit(beta[1] + beta[2]*s1 + beta[3]*i1 + beta[4]*passengerClass + beta[5]*ps1*s1) - invlogi
```

```
## Warning in Ops.factor(beta[4], passengerClass): '*' not meaningful for factors

## Warning in Ops.factor(beta[3], idade): '*' not meaningful for factors

## Warning in Ops.factor(beta[4], passengerClass): '*' not meaningful for factors

## Warning in Ops.factor(beta[5], passengerClass): '*' not meaningful for factors

## Warning in Ops.factor(beta[5] * passengerClass, sexo): '*' not meaningful for
## factors
```

```
mean(prob)
```

```
## [1] NA
```

```
#prob <- invlogit(beta[1] + beta[2]*sexo + beta[3]*idade + beta[4]*passengerClass + #beta[5]*passengerC

#probFactor <- invlogit(beta[1] + beta[2]*s1 + beta[3]*i1 + beta[4]*idade + #beta[5]*passengerClass + b
#mean(probFactor)
```

Eu deixo aqui o script da minha tentativa de montar cenários. Na hora de escrever o comando de probabilidade, eu não soube diferenciar os fatores e seus respectivos coeficientes, e o R chamou minha atenção para que esta operação não pudesse ser feita em Factors. Eu não soube utilizar os factor com a interação para calcular diferentes probabilidades.

9

```
wells <- read.table("wells.dat.txt")

glimpse(wells)
```

```
## Rows: 3,020
## Columns: 5
## $ switch  <int> 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ arsenic <dbl> 2.36, 0.71, 2.07, 1.15, 1.10, 3.90, 2.97, 3.24, 3.28, 2.52, 3.~
## $ dist    <dbl> 16.826, 47.322, 20.967, 21.486, 40.874, 69.518, 80.711, 55.146~
## $ assoc   <int> 0, 0, 0, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1,~
## $ educ    <int> 0, 0, 10, 12, 14, 9, 4, 10, 0, 0, 5, 0, 0, 0, 0, 7, 7, 7, 0, 1~
```

```
wells$switch <- as.factor(wells$switch)
wells$dist100 <- wells$dist/100  # para facilitar a interpretação

xtabs(~ switch, data = wells)
```

```
## switch
##    0    1
## 1283 1737
```

```
stargazer(wells, type = "text")
```

```
##
## ===============================================================
## Statistic   N    Mean  St. Dev.  Min  Pctl(25) Pctl(75)   Max
## ---------------------------------------------------------------
## arsenic   3,020 1.657   1.107   0.510  0.820    2.200    9.650
## dist      3,020 48.332  38.479  0.387  21.117   64.041  339.531
## assoc     3,020 0.423   0.494     0      0        1        1
## educ      3,020 4.828   4.017     0      0        8       17
## dist100   3,020 0.483   0.385   0.004  0.211    0.640    3.395
## ---------------------------------------------------------------
```

9.a Rodanda as regressões:

```
fit1_9 <- glm(switch ~ dist100 + log(arsenic) + log(arsenic):dist100, data= wells, family = "binomial")
stargazer(fit1_9, type = "text")
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                 switch
## ------------------------------------------------
## dist100                        -0.874***
##                                 (0.134)
##
## log(arsenic)                   0.983***
##                                 (0.110)
##
## dist100:log(arsenic)            -0.231
##                                 (0.183)
##
## Constant                       0.491***
##                                 (0.068)
##
## ------------------------------------------------
## Observations                     3,020
## Log Likelihood                -1,948.387
## Akaike Inf. Crit.              3,904.775
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Observando a significância das variáveis. O log de arsênico apresentou sinal positivo, indicando que quanto maior o nível de arsênico, acrescentada uma unidade a esta variável, maior a probabilidade da pessoa mudar de poço. Em razão do sinal negativo, temos que o que diminui a probabilidade de um indivíduo mudar de poço é a distância e a interação entre a distância e log da concentação de arsênico. Isso quer dizer que se o poço seguro for muito longe, menor é a chance da pessoa buscar essa fonte. E a interação, nos diz que, mesmo com um aumento na concentração do químico, se o poço mais seguro for muito distânte, menor será a probalidade de mudança. Contudo, esta última foi a única a não apresentar significância estatística.

```
invlogit(coef(fit1_9))
```

```
##          (Intercept)              dist100         log(arsenic)
##            0.6204244            0.2945258            0.7277851
## dist100:log(arsenic)
##            0.4425269
```

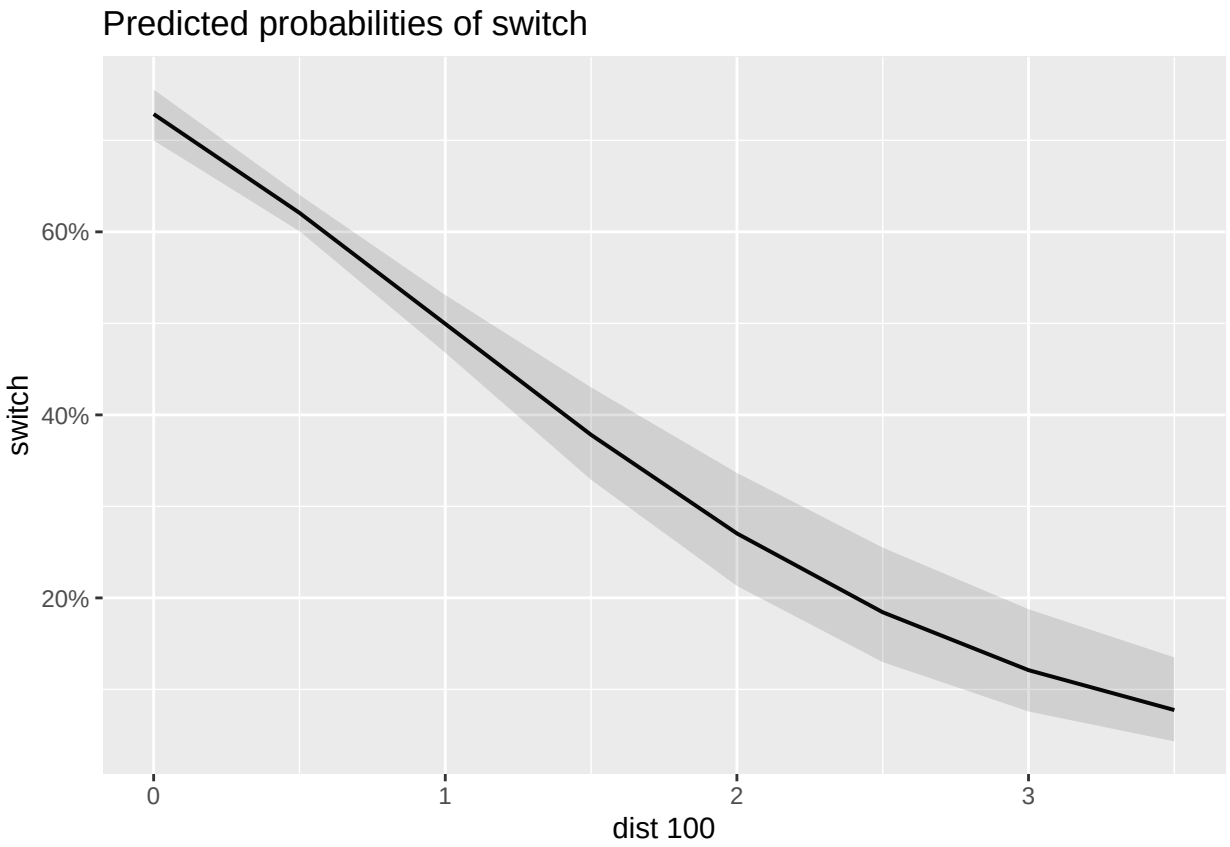
```
sjPlot::plot_model(fit1_9, type = "pred")
```

9.b

```
## Data were 'prettified'. Consider using `terms="dist100 [all]"` to get smooth plots.
```

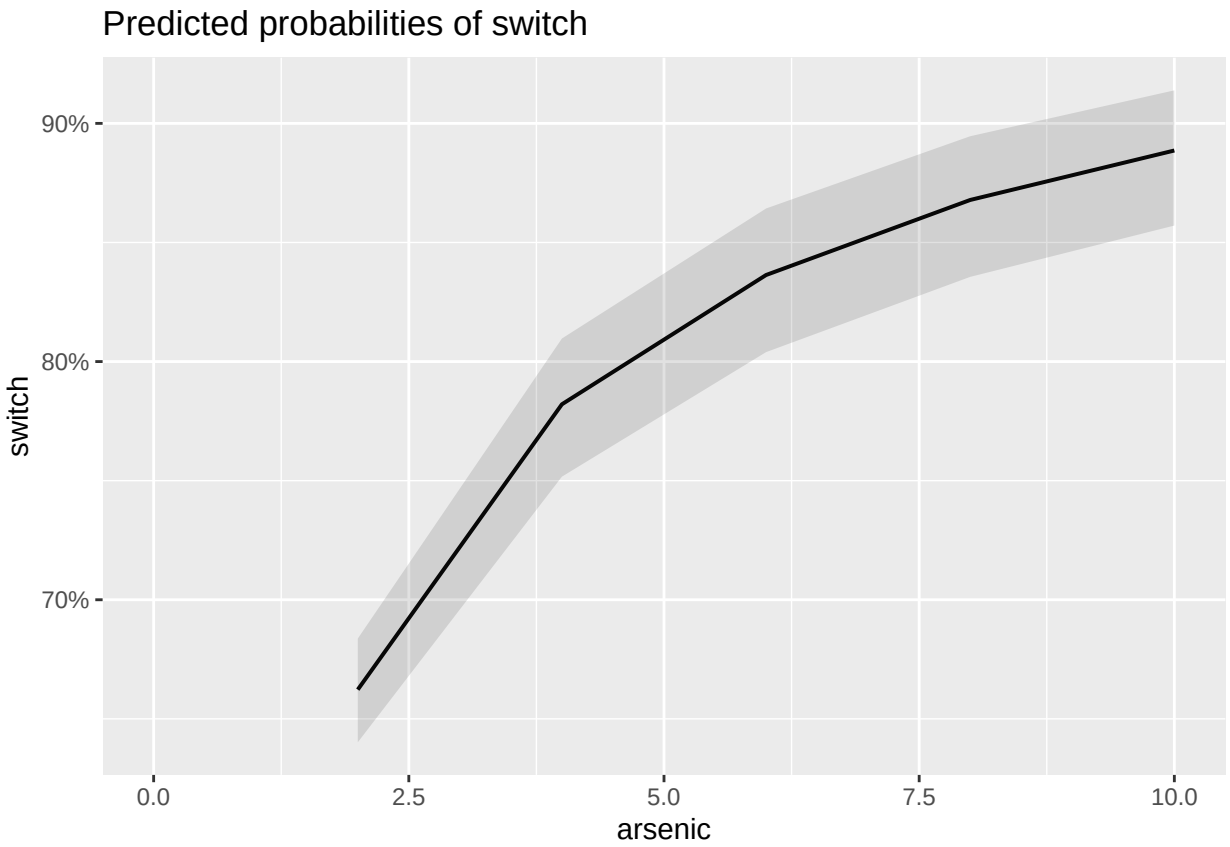
```
## Data were 'prettified'. Consider using `terms="arsenic [all]"` to get smooth plots.
```

```
## $dist100
```




```
##
## $arsenic
```

```
## Warning: Removed 1 row(s) containing missing values (geom_path).
```



Podemos observar nos gráficos acima, que: Para a concentração de arsênico, a medida que cresce, aumenta a probabilidade de um indíduo se mudar para um poço mais seguro. O inverso acontece com a distância. Quanto maior esta for, menor a probabilidade de trocar.

9.c Sabendo que a interpretação do coeficente depende do valor da variável independente utilizada, ou seja de X, realizaremos a leitura dada as seguintes comparações:

- Acréscimo em 100 metros de distância, com arsênico constante:

```
attach(wells)
beta <- coef(fit1_9)
hi <- 1
lo <- 0

prob <- invlogit(beta[1] + beta[2]*hi + beta[3]*log(arsenic) + beta[4]*log(arsenic)*hi)- invlogit(beta[
mean(prob)
```

```
## [1] -0.2113356
```

O aumento em 100 metros diminui 21%, em média, a probabilida de mudança a um poço mais seguro

- Diferença entre 100 e 200 metros, mantendo arsenico constante:

```
beta <- coef(fit1_9)
hi <- 2
lo <- 1

prob <- invlogit(beta[1] + beta[2]*hi + beta[3]*log(arsenic) + beta[4]*log(arsenic)*hi)- invlogit(beta[
mean(prob)
```

```
## [1] -0.2090207
```

```
A diferença entre 100 metros a mais diminui em 20% a mudança de poço. A aumento na distância causou pou
```

- Diferença na concentração entre 0,5 e 1,0 de arsênico, mantendo distância constante

```
beta <- coef(fit1_9)
hi <- 1
lo <- 0.5

prob <- invlogit(beta[1] + beta[2]*dist100 + beta[3]*log(hi) + beta[4]*dist100*log(hi))- invlogit(beta[
mean(prob)
```

```
## [1] 0.1460174
```

O aumento em 0.5 de arsênico tem um impacto de 14% na probabilidade de mudança, em média, com distância constante.

- Diferença na concentração entre 1,0 e 2,0 de arsênico, mantendo distância constante:

```
beta <- coef(fit1_9)
hi <- 2
lo <- 1

prob <- invlogit(beta[1] + beta[2]*dist100 + beta[3]*log(hi) + beta[4]*dist100*log(hi))- invlogit(beta[
mean(prob)
```

```
## [1] 0.1404344
```

```
detach(wells)
```

O aumento de uma unidade de concentração de arsênico é o mesmo de 0.5, 14%. Parece que esta quantia não afeta tanto a decisão de mudança ou não de poço, nestes cenários.

10

```
library(car)
data("Cowles")
head(Cowles)
```

```
##   neuroticism extraversion    sex volunteer
## 1          16           13 female        no
## 2           8           14   male        no
## 3           5           16   male        no
## 4           8           20 female        no
## 5           9           19   male        no
## 6           6           15   male        no
```

```
summary(Cowles)
```

```
##   neuroticism     extraversion       sex      volunteer
##  Min.   : 0.00   Min.   : 2.00   female:780   no :824
##  1st Qu.: 8.00   1st Qu.:10.00   male  :641   yes:597
##  Median :11.00   Median :13.00
##  Mean   :11.47   Mean   :12.37
##  3rd Qu.:15.00   3rd Qu.:15.00
##  Max.   :24.00   Max.   :23.00
```

```
glimpse(Cowles)
```

```
## Rows: 1,421
## Columns: 4
## $ neuroticism  <int> 16, 8, 5, 8, 9, 6, 8, 12, 15, 18, 12, 9, 13, 9, 12, 11, 5~
## $ extraversion <int> 13, 14, 16, 20, 19, 15, 10, 11, 16, 7, 16, 15, 11, 13, 16~
## $ sex          <fct> female, male, male, female, male, male, female, male, mal~
## $ volunteer    <fct> no, no, no, no, no, no, no, no, no, no, no, no, no, no, n~
```

```
Cowles <- Cowles %>%
  mutate(voluntarioNum = case_when( volunteer == "yes" ~ 1,
                                    volunteer == "no" ~ 0))
Cowles$voluntarioNum <- as.factor(Cowles$voluntarioNum)
```

10.a O esperado da pesquisa é que se voluntariar ou não (var. dicotômica) dependendesse do sexo do indivíduo somado interação entre fatores de sua personalidade(extroversão e neuroticismo). Sendo assim, temos o seguinte modelo

```
fit1_10 <- glm(voluntarioNum ~ sex + neuroticism + extraversion + neuroticism:extraversion,
               data = Cowles, family = "binomial")
stargazer(fit1_10, type= "text")
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                             voluntarioNum
## ------------------------------------------------
## sexmale                       -0.247**
##                                (0.112)
##
## neuroticism                   0.111***
##                                (0.038)
```

```
##
## extraversion                       0.167***
##                                     (0.038)
##
## neuroticism:extraversion           -0.009***
##                                     (0.003)
##
## Constant                           -2.358***
##                                     (0.501)
##
## ---------------------------------------------------
## Observations                         1,421
## Log Likelihood                     -948.720
## Akaike Inf. Crit.                  1,907.440
## ===================================================
## Note:                     *p<0.1; **p<0.05; ***p<0.01
```

10.b Do ponto de vista da significância, todas as variáveis a apresentaram. A com menor magnitude e significância foi a interação entre neuroceticistmo e extroversão. Como os pesquisadores imaginaram, sexo importa para o voluntariado. Ser homem possui, aqui, um efeito negativo sobre a participação. Tal qual a interação do modelo. Porém, quando isoladas, tanto o neuroticismo quanto a extroversão apresentaram um efeito positivo, e bem baixo coeficiente de erro padrão. Sobre a constante, ela apresenta signficância e sinal negativo.

```
# Taxa de erro

prev <- predict(fit1_10, type = "response")
erro <- mean((prev > 0.5 & Cowles$voluntarioNum == 0) | (prev < .5 & Cowles$voluntarioNum == 1))
erro
```

10.c

```
## [1] 0.4081633
```

A taxa de erro do modelo deu um valor alto de 40,18%.

10.d Centralizando as variáveis de neuroticismo e extroversão, nós podemos interpretar a interação com mais facilidade:

```
Cowles <- Cowles %>%
  mutate(c.neuro = neuroticism - mean(neuroticism),
         c.extrav = extraversion - mean(extraversion))
fit2_10 <- glm(voluntarioNum ~ sex + c.neuro + c.extrav + c.neuro:c.extrav,
               data = Cowles, family = "binomial")

stargazer(fit2_10, type = "text")
```

```
##
## ===============================================
```

```
##                         Dependent variable:
##                     ---------------------------
##                               voluntarioNum
## ------------------------------------------------
## sexmale                        -0.247**
##                                 (0.112)
##
## c.neuro                         0.005
##                                 (0.011)
##
## c.extrav                       0.069***
##                                 (0.014)
##
## c.neuro:c.extrav               -0.009***
##                                 (0.003)
##
## Constant                       -0.237***
##                                 (0.074)
##
## ------------------------------------------------
## Observations                    1,421
## Log Likelihood                 -948.720
## Akaike Inf. Crit.              1,907.440
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

```
coefficients(fit2_10)
```

```
##      (Intercept)          sexmale          c.neuro          c.extrav
##     -0.237331701     -0.247152026      0.004957182      0.068718908
## c.neuro:c.extrav
##     -0.008552465
```

```
tibble(Variavel = c("intercepto", "sexmale", "c.neuro", 'c.extrav', "c.neuro:c.extrav"),
       beta_4 = c(coef(fit2_10)[1]/4, coef(fit2_10)[2]/4, coef(fit2_10)[3]/4,
                  coef(fit2_10)[4]/4, coef(fit2_10)[5])) %>%
  qflextable()
```

```
## Warning: Warning: fonts used in ‘flextable‘ are ignored because the ‘pdflatex‘
## engine is used and not ‘xelatex‘ or ‘lualatex‘. You can avoid this warning
## by using the ‘set_flextable_defaults(fonts_ignore=TRUE)‘ command or use a
## compatible engine by defining ‘latex_engine: xelatex‘ in the YAML header of the
## R Markdown document.
```

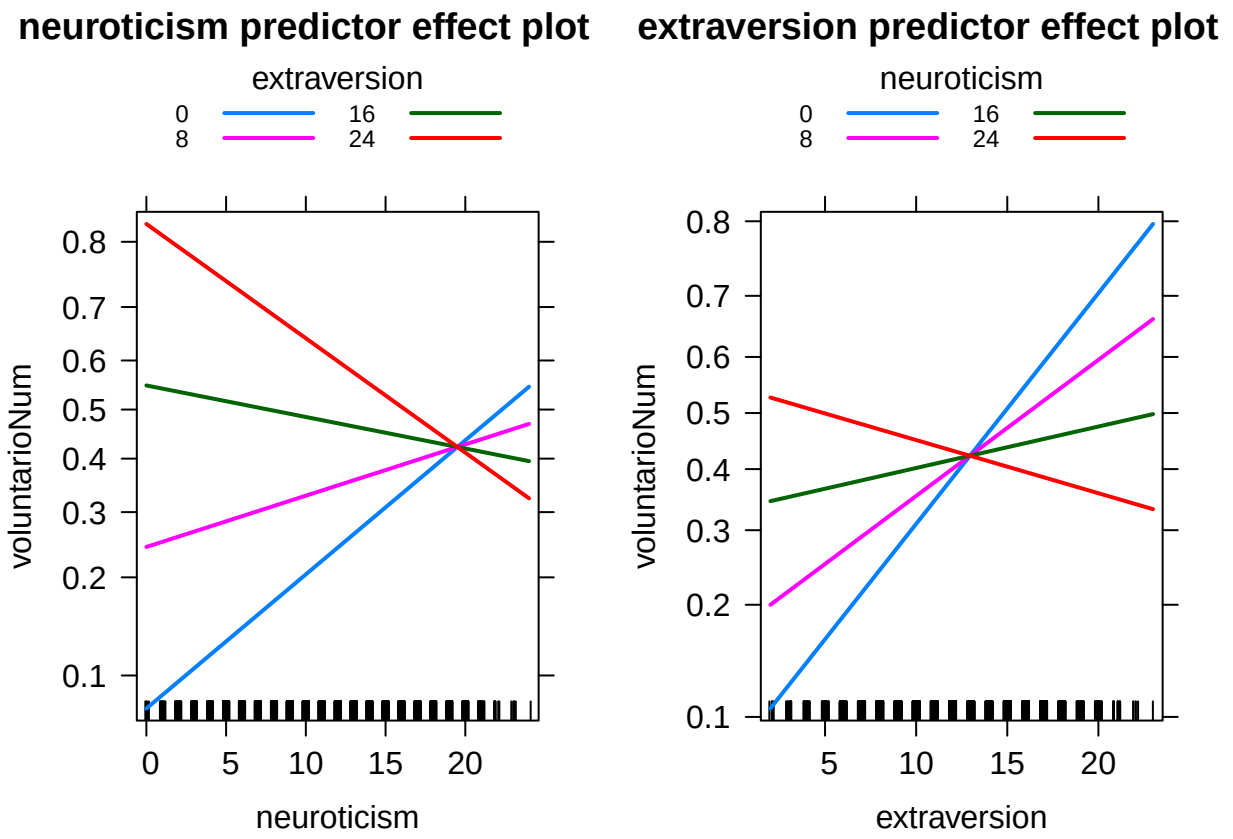
Variavel	beta_4
intercepto	-0.059332925
sexmale	-0.061788006
c.neuro	0.001239296
c.extrav	0.017179727
c.neuro:c.extrav	-0.008552465

A interpretação dos efeitos: houve uma perda de efeitos em todas as variáveis. Neurotismo perdeu sua significância estatística. Os sinais se mantiveram os mesmo

Sobre a interpretação dos coeficientes: Uma vez centralizado em torno da média, podemos de forma mais tranquila interpretar os coeficentes do modelo. A probalidade de um indivíduo, quanto todas as outras variáveis estão na média é de 44%. O impacto de uma unidade de neuroticismo representa 0,12% no aumento da participação quando quando a variável extroversão estiver na média. Quando o neurocetismo estiver na média, o impacto do aumento de uma unidade em extroversão é de 0.17%. Essas duas probabilidades nos indicam que a interação é muito baixa, as duas variáveis são independentes, uma da outra.

10.e plot(predictorEffects(cwl.glm, ~neuroticism + extraversion, xlevels = list(neuroticism=seq(0,24, by = 8), extraversion=seq(0,24, by = 8))), lines = list(multiline = T))

```
plot(predictorEffects(fit1_10, ~neuroticism + extraversion, xlevels = list(neuroticism=seq(0,24, by = 8
```



- Primeiro plot, onde eixo-x é neuroticismo:

  Quando a extroversão é zero, se acrecentarmos uma unidade em neuroticismo, a probabilidade de se voluntariar tende a crescer. A diferença entre 0 e 20 unidades de neuroticismo é bem grande. Com extrov =0, quando neuro = 0, a prob de se voluntariar é quase zero. Porém, em neuro>20, a prob chega a 60%. Quando extroversão = 8, a probabilidade permanece crescente, mas a diferença entre um acréscimo muito ou pouco de neurocetismo é baixo. Porém o sinal da reta se inverte conforme aumentamos as unidades de extroversão. A partir de extroversão=16, a probabilidade de ser voluntariado diminui conforme aumentamos as unidades em neuroticismo. O mesmo ocorre com extroversão=24, porém mais acentuado. Há uma queda de 50% na probabilidade de se voluntariar.

Todas as retas de probabilidade, com valores diferentes de extroversão, se encontram em um valor próximo a 20u. de neuroticismo. É possível observar como a probabilidade se modifica com o acréscimo nas duas variáveis.

- Segundo plot, quando eixo-x é extrversão:

  Quando neuroticismo = 0, quanto maior o acréscimo em unidades em extroversão, maior é a probabilidade de se voluntariar. Quando extraversão > 20u., a prob. atinge quase 80%. Porém, observamos que, ao aumentarmos as unidades de neuroticismo, o sinal do impacto se torna negativo. Isto é, quando neuro=24, há um queda de 0.2 na probabilidade de se voluntariar.

  Novamente, as retas de probabilidade, se econtram em um valor entre 10 e 15u. de extroversão.

  O que podemos concluir nesta análise é que neuroticismo e extroversão possuem interação. Isto é visível pelo comportamento das retas de probabilidades, como elas se alteram conforme o incremento de unidades de neuroticismo e extroversão.