Regressão Logística

Bioestatística em R

André M Ribeiro-dos-Santos

13 de Abr, 2017

Regressão Logística

Objetivos

- Modelar a relação entre variáveis a uma variável binomial
- Compreender os significado dos coeficientes.
- Ilustrar uma regressão logística e o efeito dos coeficientes.

Imagine

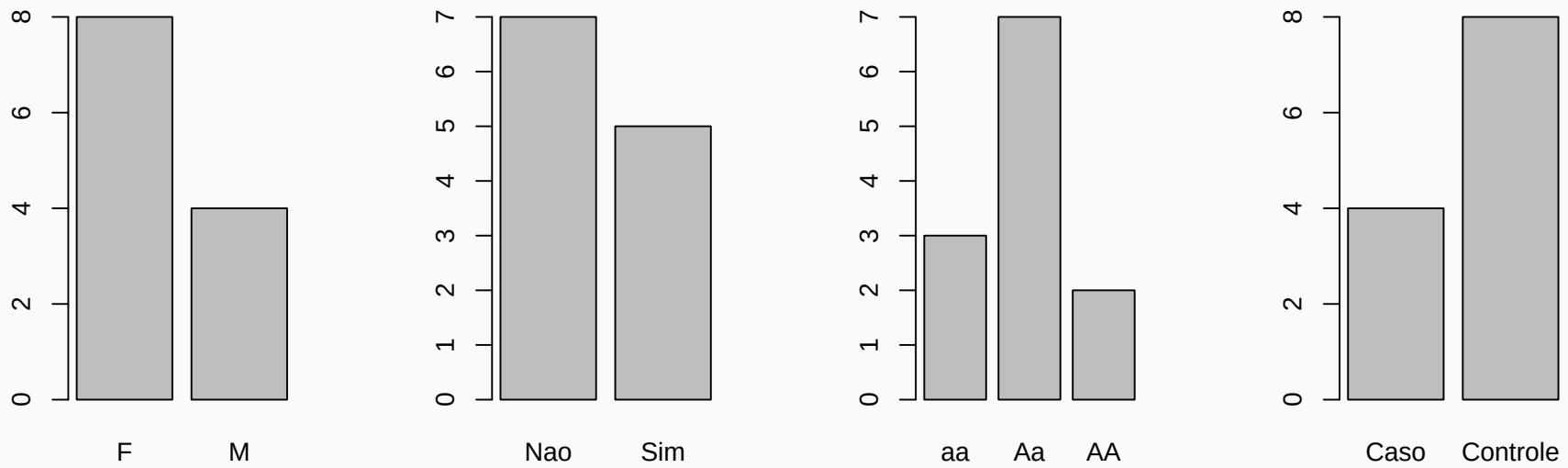
Em um estudo caso/controle, pesquisadores desejavam associar a presença de câncer com um polimorfismo do gene CDH1, no entanto fatores como sexo e tabagismo influenciam o risco do do câncer.

Qual o risco de desenvolver câncer em função da presença do alelo mutante?

Table 1: Dados de estudo caso/controle para câncer

sexo	tabagismo	CDH1	Cancer	sexo	tabagismo	CDH1	Cancer
M	Nao	Aa	Controle	F	Nao	aa	Controle
F	Sim	AA	Controle	F	Nao	Aa	Caso
F	Nao	Aa	Caso	F	Nao	Aa	Caso
F	Sim	aa	Controle	M	Sim	Aa	Controle
M	Sim	Aa	Controle	M	Nao	Aa	Controle
F	Sim	aa	Controle	F	Nao	AA	Caso

```
> sex <- c("M", "F", "F", "F", "M", "F", "F", "F", "F", "M", "M", "F")
> tab <- c("Nao", "Sim", "Nao", "Sim", "Sim", "Sim",
+          "Nao", "Nao", "Nao", "Sim", "Nao", "Nao")
> cdh1 <- c("Aa", "AA", "Aa", "aa", "Aa", "aa",
+           "aa", "Aa", "Aa", "Aa", "Aa", "AA")
> cancer <- c("Controle", "Controle", "Caso", "Controle", "Controle",
+             "Controle", "Controle", "Caso", "Caso", "Controle",
+             "Controle", "Caso")
> par(mfrow = c(1,4))
> barplot(table(sex));  barplot(table(tab))
> barplot(table(cdh1)); barplot(table(cancer))
```



Avaliando o problema

- Qual o tipo das variáveis sendo relacionadas?
- Qual o objetivo?

Avaliando o problema

- Qual o tipo das variáveis sendo relacionadas? **Queremos associar duas variáveis categóricas, sendo uma binomial**
- Qual o objetivo? **Queremos associar duas variáveis categóricas, medindo a influência de uma sobre a outra**

Se quiséssemos somente associar o genótipo a presença ou não da doença usaríamos um **qui-quadrado**.

```
> chisq.test(table(cancer, cdh1))
```

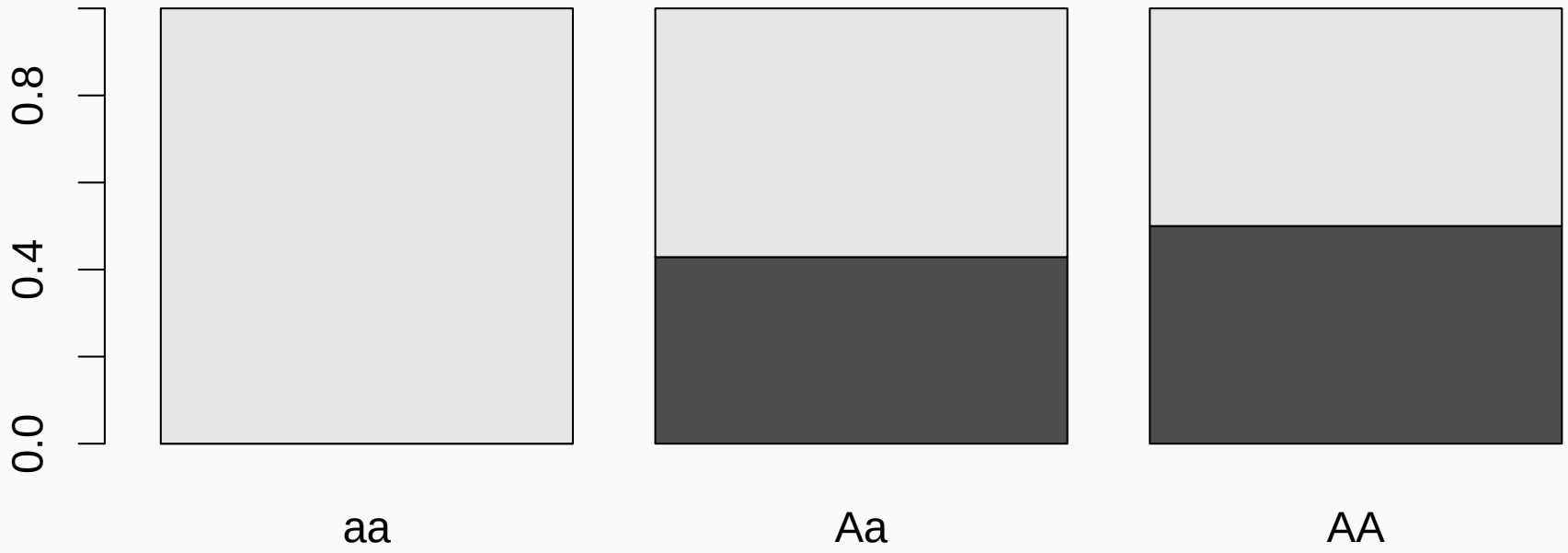
```
##
##  Pearson's Chi-squared test
##
## data:  table(cancer, cdh1)
## X-squared = 2.0357, df = 2, p-value = 0.3614
```

No entanto, essa análise não indica a intensidade da relação, nem é capaz de considerar **confundidores** (Outros fatores que sabe-se estar relacionado com a resposta, mas não é o alvo principal).

Regressão Logística

Para esses casos, a análise mais indicada é uma **regressão logística**. Esta análise é uma especificação de um regressão linear que no lugar de buscar relacionar medidas a um variável resposta quantitativa, ela relaciona a uma variável resposta binária (**TRUE** ou **FALSE**).

```
> barplot(prop.table(table(cancer, cdh1), 2))
```



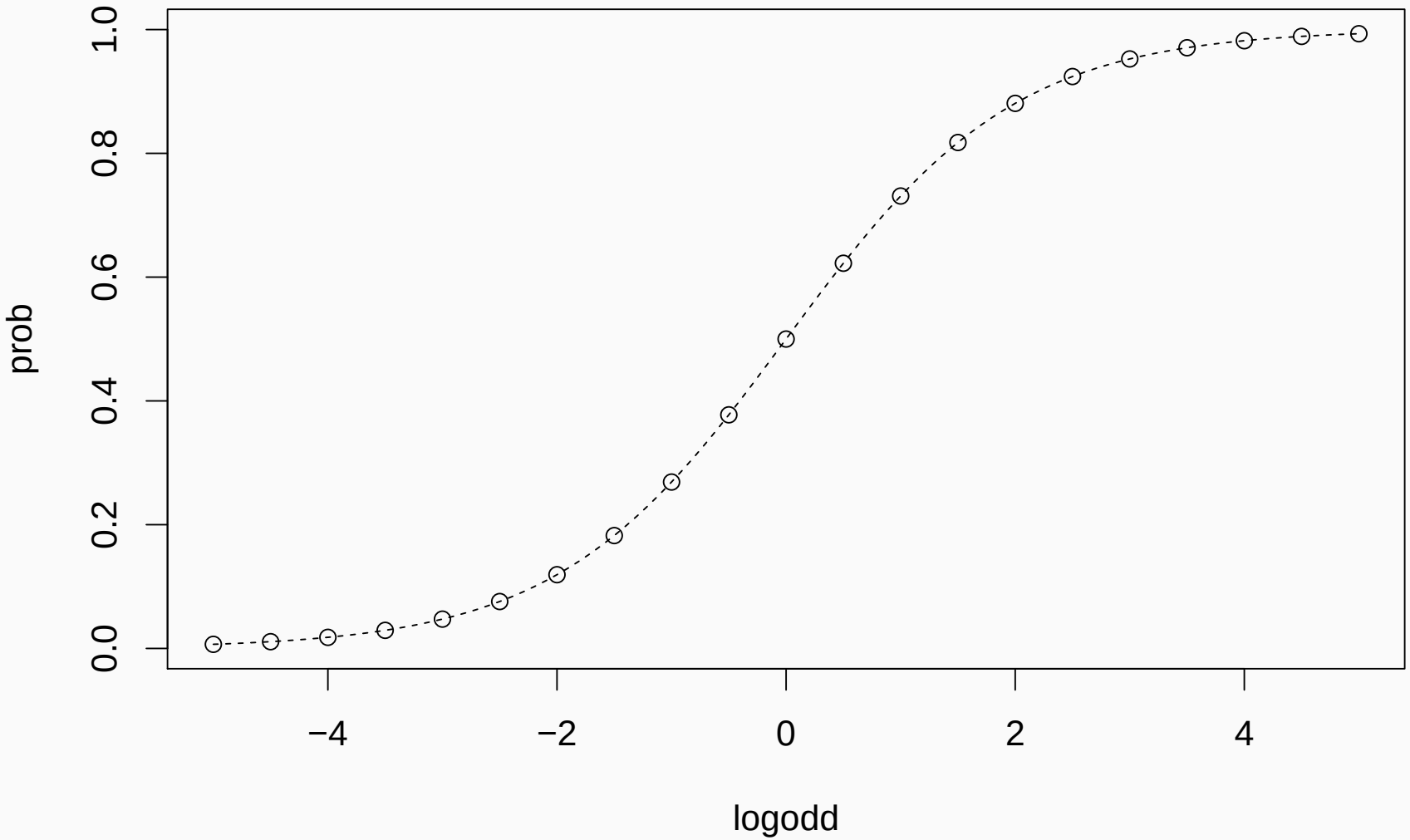
Isso é possível pela conversão das probabilidades na escala linear do **log-odds**.

$$logit(p_i) = ln(\frac{p_i}{1 - p_i}) = \alpha + \beta * x + \epsilon$$

ou

$$p_i = \frac{1}{1 + e^{-(\alpha + \beta * x + \epsilon)}}$$

```
> logodd <- seq(-5, 5, 0.5)
> prob   <- 1 / (1 + exp(-logodd)) ## ou plogis(logodd)
> plot(logodd, prob)
> curve(plogis, lty=2, add = TRUE)
```



```
> ?glm
> ## glm                    package:stats                R Documentation
> ## Fitting Generalized Linear Models
> ## Description:
> ##      'glm' is used to fit generalized linear models, specified by
> ##      giving a symbolic description of the linear predictor and a
> ##      description of the error distribution.
> ## Usage:
> ##      glm(formula, family = gaussian, data, weights, subset,
> ##          na.action, start = NULL, etastart, mustart, offset,
> ##          control = list(...), model = TRUE, method = "glm.fit",
> ##          x = FALSE, y = TRUE, contrasts = NULL, ...)
```

```
> has_cancer = (cancer == "Caso")
> (model <- glm(has_cancer~cdh1, family = "binomial"))


##
## Call:  glm(formula = has_cancer ~ cdh1, family = "binomial")
##
## Coefficients:
## (Intercept)       cdh1Aa       cdh1AA
##      -18.57        18.28        18.57
##
## Degrees of Freedom: 11 Total (i.e. Null);  9 Residual
## Null Deviance:       15.28
## Residual Deviance: 12.33     AIC: 18.33
```

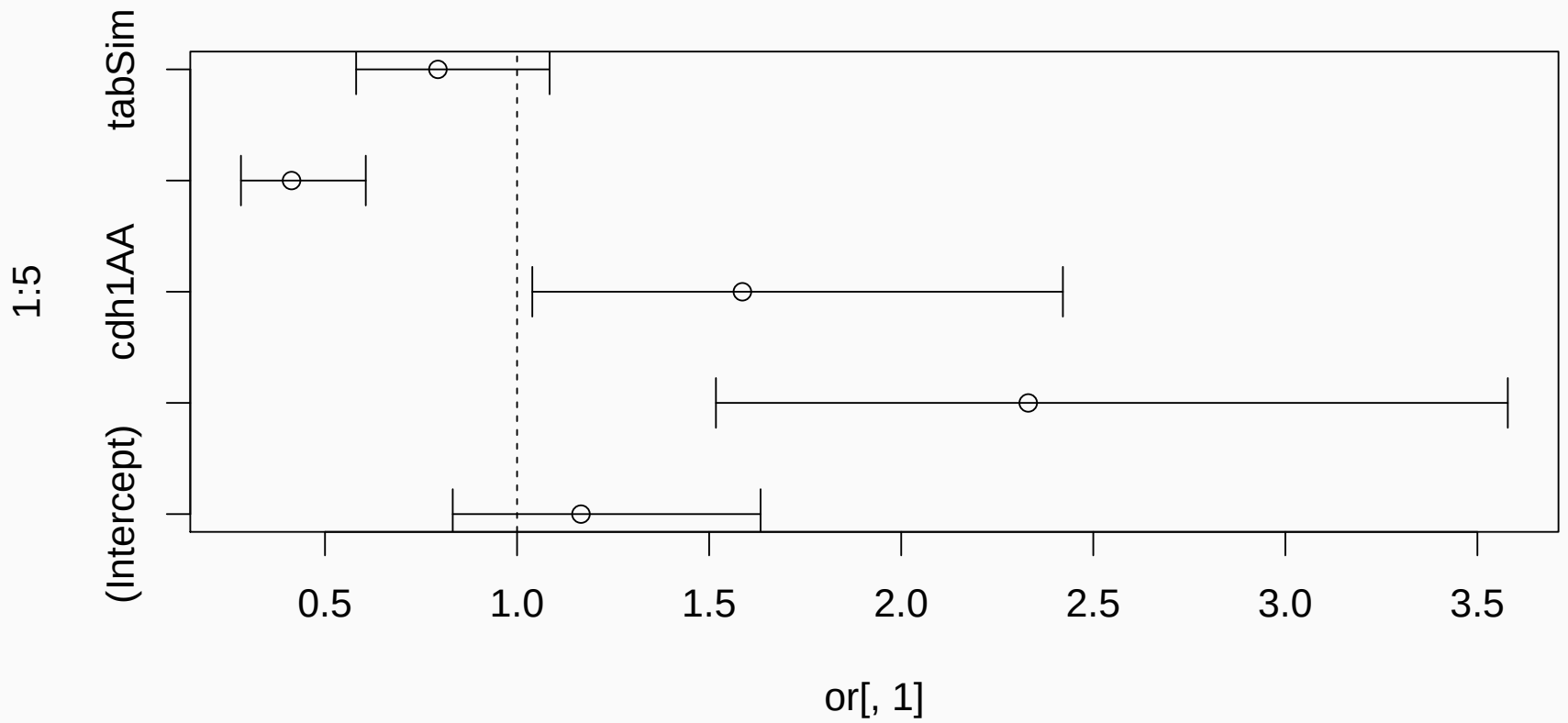
```
> summary(model)
```

```
##
## Call:
## glm(formula = has_cancer ~ cdh1, family = "binomial")
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.17741  -1.05794  -0.00013   1.20850   1.30177
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -18.57    3765.85  -0.005    0.996
## cdh1Aa         18.28    3765.85   0.005    0.996
## cdh1AA         18.57    3765.85   0.005    0.996
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 15.276  on 11  degrees of freedom
## Residual deviance: 12.333  on  9  degrees of freedom
## AIC: 18.333
##
## Number of Fisher Scoring iterations: 17
```

```
> ## Log-odds effect
> (eff <- cbind(coef(model), confint(model)))


## Waiting for profiling to be done...


##                            2.5 %  97.5 %
## (Intercept) -18.56607         NA 472.844
## cdh1Aa       18.27839 -374.0779      NA
## cdh1AA       18.56607 -755.8809      NA

> ## Odds-Ratio effect
> exp(eff)


##                                  2.5 %        97.5 %
## (Intercept) 8.646869e-09             NA 2.256982e+205
## cdh1Aa      8.673659e+07 3.467725e-163            NA
## cdh1AA      1.156488e+08  0.000000e+00            NA
```

```
> model <- glm(has_cancer~cdh1+sex+tab)
> summary(model)


##
## Call:
## glm(formula = has_cancer ~ cdh1 + sex + tab)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.38462  -0.11538   0.00000   0.08654   0.38462
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.1538     0.1720   0.894  0.40080
## cdh1Aa        0.8462     0.2189   3.866  0.00616 **
## cdh1AA        0.4615     0.2156   2.141  0.06958 .
## sexM         -0.8846     0.1959  -4.515  0.00275 **
## tabSim       -0.2308     0.1592  -1.449  0.19057
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.05494505)
##
##     Null deviance: 2.66667  on 11  degrees of freedom
```

```
> or <- exp(cbind(coef(model), confint(model)))

## Waiting for profiling to be done...

> plot(or[,1], 1:5, xlim = range(or), yaxt= "n")
> arrows(or[,2], 1:5, or[,3], 1:5, angle = 90, length = 0.15, code=3)
> abline(v = 1, lty = 2)
> axis(2, 1:5, rownames(or))
```



Bonus - Modelos Genéticos

Additivo	Recessivo	Dominante
aa < Aa < AA	aa/Aa vs AA	aa vs Aa/AA

```
> add <- as.numeric(factor(cdh1))
> dom <- cdh1 != "aa"
> rec <- cdh1 == "AA"
```

Exercícios - Regressão Logística

Em um estudo de cancer de mama, os pesquisadores desejam avaliar a associação de dois polimorfismos com o risco de desenvolver leucemia. Sobre o *dataset* responda:

1. Dado todas as medidas avaliadas (`sex`, `cigar`, `amr`, `eur`, e `afr`), avalie quais apresentam diferenças entre os casos e controle.
2. Ilustre a diferença de frequencia de casos e controle entre os alelos de ambos polimorfismos invesitgados (`cdh1`, `tp53`). Algum modelo de efeito genético parece mais plausível?
3. Utilizando as medidas com diferença entre caso e controle como confundidores, modele o risco de desenvolver leucemia em função de cada polimorfismo. Utilize o modelo genético que considerar mais adequado.
4. Ilustre o odds-ratio do polimorfismo para os modelos desenvolvidos na questão anterior.

```
> cancer <- read.table('cancer-leucemia.tsv', header=T)
```

1. Dado todas as medidas avaliadas (`sex`, `cigar`, `amr`, `eur`, e `afr`), avalie quais apresentam diferenças entre os casos e controle.

```
> fisher.test(table(cancer$sex, cancer$cancer))$p.value
```

```
## [1] 0.01477669
```

```
> fisher.test(table(cancer$cigar, cancer$cancer))$p.value
```

```
## [1] 0.03858585
```

```
> wilcox.test(amr~cancer, data=cancer)$p.value
```

```
## [1] 0.6546313
```

```
> wilcox.test(afr~cancer, data=cancer)$p.value
```

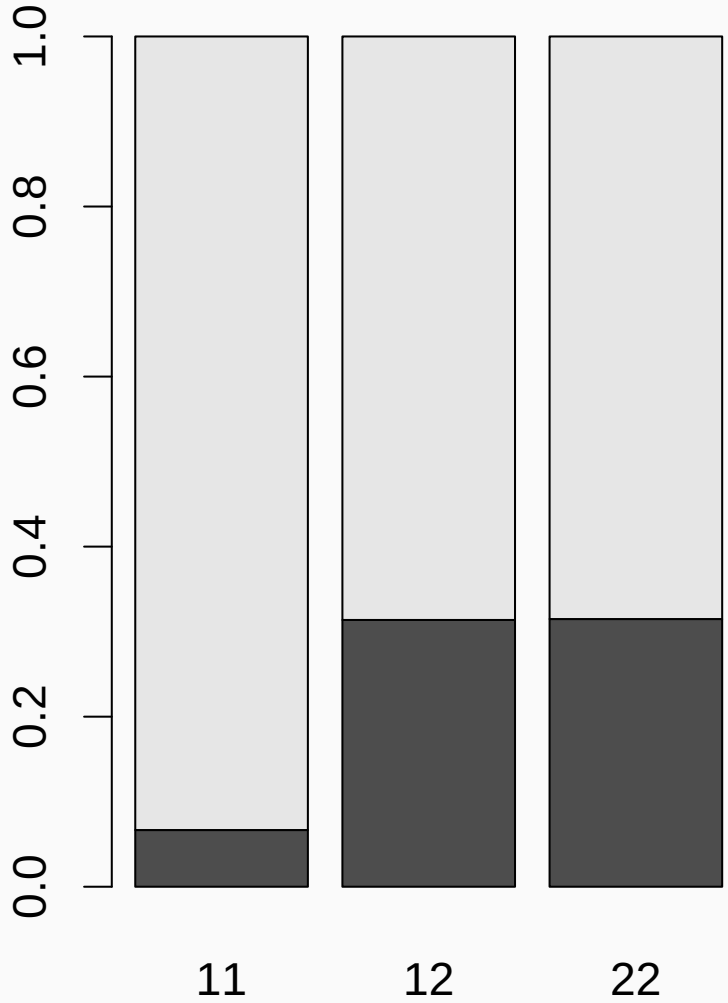
```
## [1] 0.3623085
```

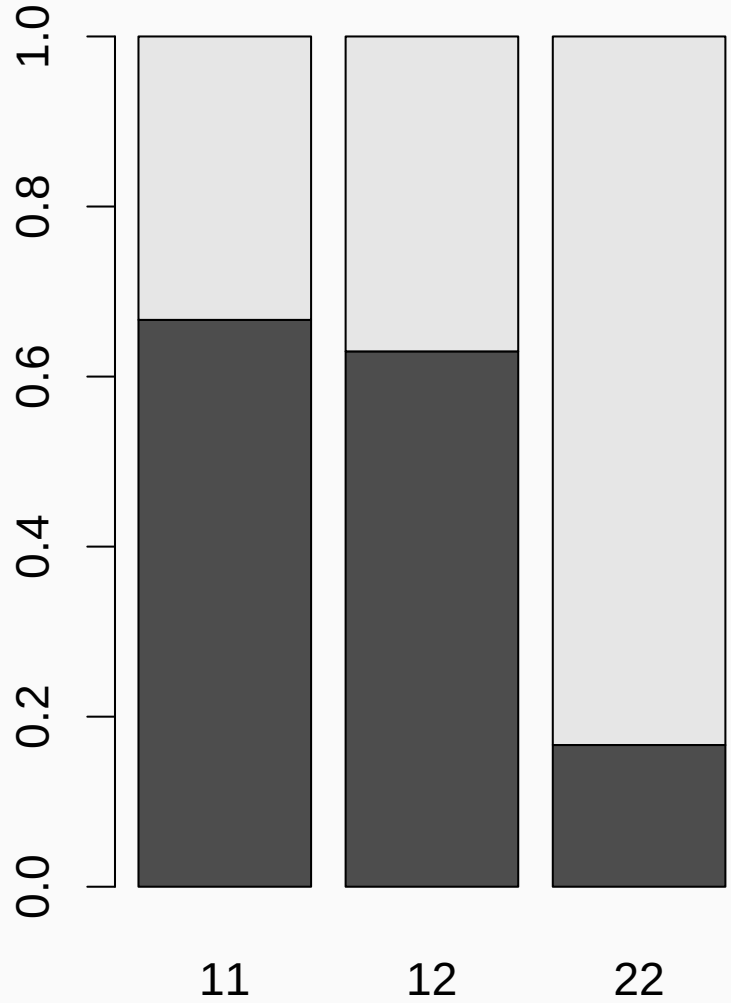
```
> wilcox.test(eur~cancer, data=cancer)$p.value
```

```
## [1] 0.1476221
```

2. Ilustre a diferença de frequencia de casos e controle entre os alelos de ambos polimorfismos invesitgados (`cdh1`, `tp53`). Algum modelo de efeito genético parece mais plausível?

```
> par(mfrow = c(1,2))
> barplot(prop.table(table(cancer$cancer, cancer$cdh1),2))
> barplot(prop.table(table(cancer$cancer, cancer$tp53), 2))
```





3. Utilizando as medidas da questão 1 com confundidores, modele o risco de desenvolver leucemia em função de cada polimorfismo. Utilize o modelo genético que considerar mais adequado.

```
> cancer$has_cancer <- cancer$cancer == "Caso"
> cancer$cdh1dom <- cancer$cdh1 != "11"
> cancer$tp53rec <- cancer$tp53 == "22"
> model_cdh1 <- glm(has_cancer~cdh1dom+sex+cigar+afr+eur, data=cancer)
> model_tp53 <- glm(has_cancer~tp53rec+sex+cigar+afr+eur, data=cancer)
```

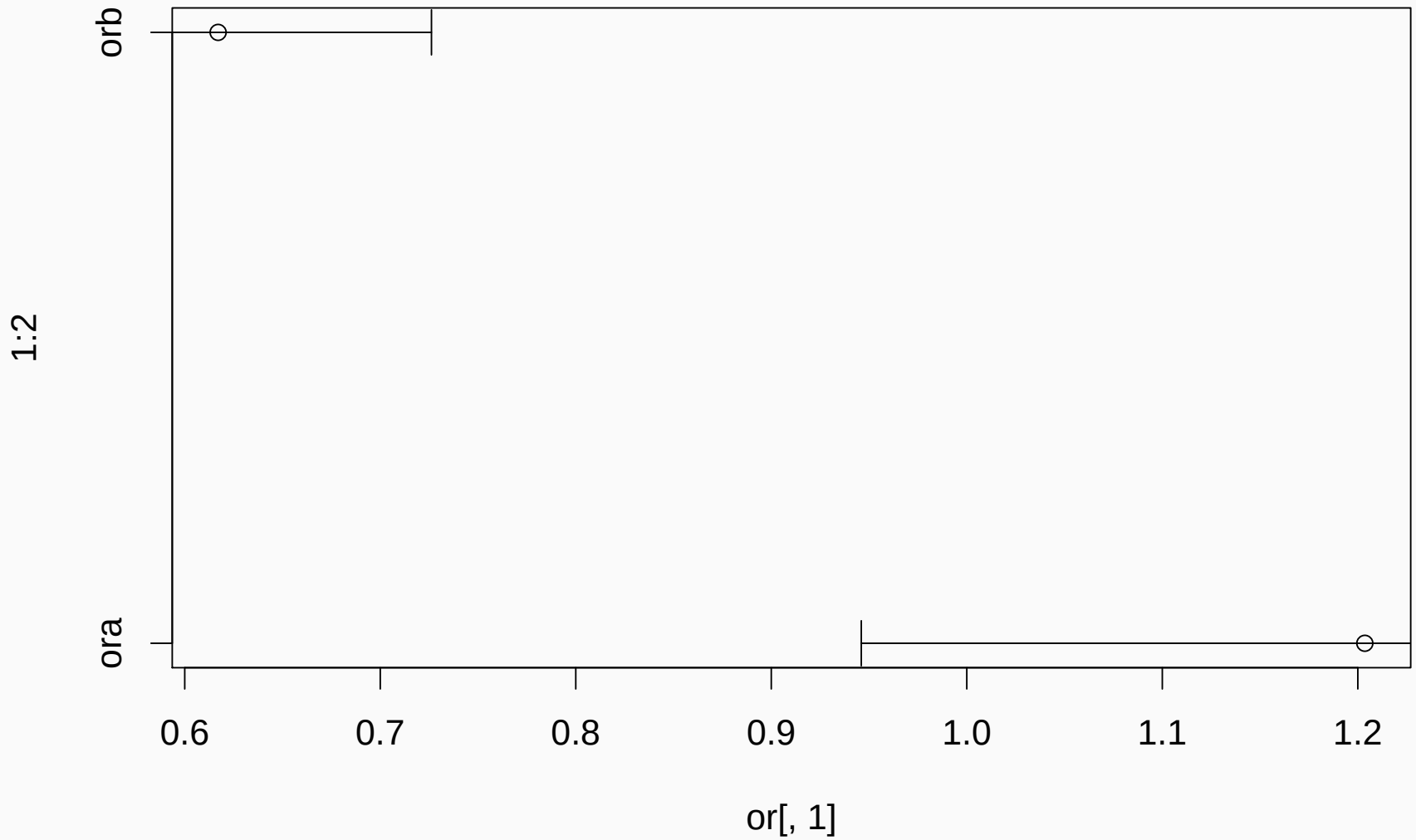
```
> summary(model_cdh1)

##
## Call:
## glm(formula = has_cancer ~ cdh1dom + sex + cigar + afr + eur,
##     data = cancer)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5530  -0.2855  -0.1839   0.4998   0.8647
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.08952    0.85768   0.104  0.91706
## cdh1domTRUE  0.18531    0.12285   1.508  0.13422
## sexM         0.21057    0.07955   2.647  0.00927 **
## cigarS      -0.28365    0.11298  -2.511  0.01345 *
## afr         -0.71499    1.17449  -0.609  0.54389
## eur          0.24289    1.06432   0.228  0.81989
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1849195)
##
```

```
> summary(model_tp53)

##
## Call:
## glm(formula = has_cancer ~ tp53rec + sex + cigar + afr + eur,
##     data = cancer)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.88209  -0.22714  -0.08817   0.22426   0.86552
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.51021    0.76836  -0.664  0.50802
## tp53recTRUE -0.48278    0.08308  -5.811 5.75e-08 ***
## sexM         0.18193    0.07047   2.582  0.01110 *
## cigarS      -0.30631    0.10030  -3.054  0.00281 **
## afr          0.42637    1.06181   0.402  0.68876
## eur          1.66592    0.96142   1.733  0.08584 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1455127)
##
```

4. Ilustre o odds-ratio do polimorfismo para os modelos desenvolvidos na questão anterior.

```
> ora <- exp(cbind(coef(model_cdh1), confint(model_cdh1)))["cdh1domTRUE",]

## Waiting for profiling to be done...

> orb <- exp(cbind(coef(model_tp53), confint(model_tp53)))["tp53recTRUE",]

## Waiting for profiling to be done...

> (or <- rbind(ora, orb))

##                   2.5 %    97.5 %
## ora 1.2035859 0.9460375 1.5312491
## orb 0.6170664 0.5243367 0.7261954
```

```
> plot(or[,1], 1:2, yaxt="n")
> axis(2, 1:2, rownames(or))
> arrows(or[,2], 1:2, or[,3], 1:2, code=3, angle=90, length=0.15)
```



Ao final...

Objetivos

- Modelar a relação entre variáveis a uma variável binomial
- Compreender os significado dos coeficientes.
- Ilustrar uma regressão logística e o efeito dos coeficientes.

Até a próxima