
# Modelos Lineares Generalizados

## unificação metodológica

Alexandre Adalardo de Oliveira
PlanECO 2017

Conceitos

- estrutura do erro
- preditora linear
- função de ligação

# Modelos Lineares Generalizados

# Função de ligação

- o preditor apresenta resposta linear $\eta$

$$\eta = \alpha + \sum \beta_i x_i$$

- a função inversa da ligação retorna à escala da preditora

$$Y = g(\eta)^{-1}$$

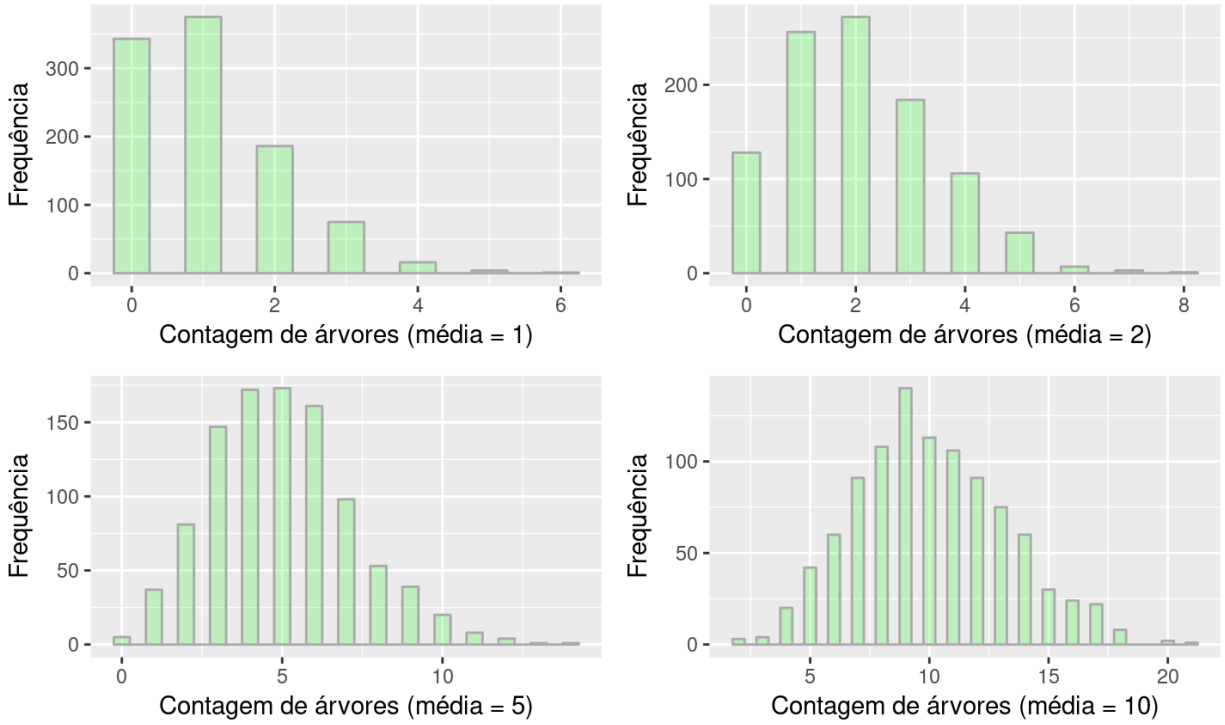
# Função de ligação: canônica

- dependendo da estrutura de erro há uma função de ligação padrão

| **resposta** | **resíduos** | **ligação** |
|---|---|---|
| **contínua** | gaussiano | identidade |
| **contagem** | poisson | log |
| **proporção** | binomial | logit |
| **binária** | binomial | logit |

# Diferentes estruturas de erro

Contagem: árvores por parcela



# Dados de Contagem

- contagem é limitada no zero
- variância não é constante (aumenta com a média)
- erro não tem distribuição normal
- valores inteiros (não contínuos)

Função de ligação

$$\log(\alpha + \sum \beta_i x_i)$$

# GLM: ajuste de modelo de contagem

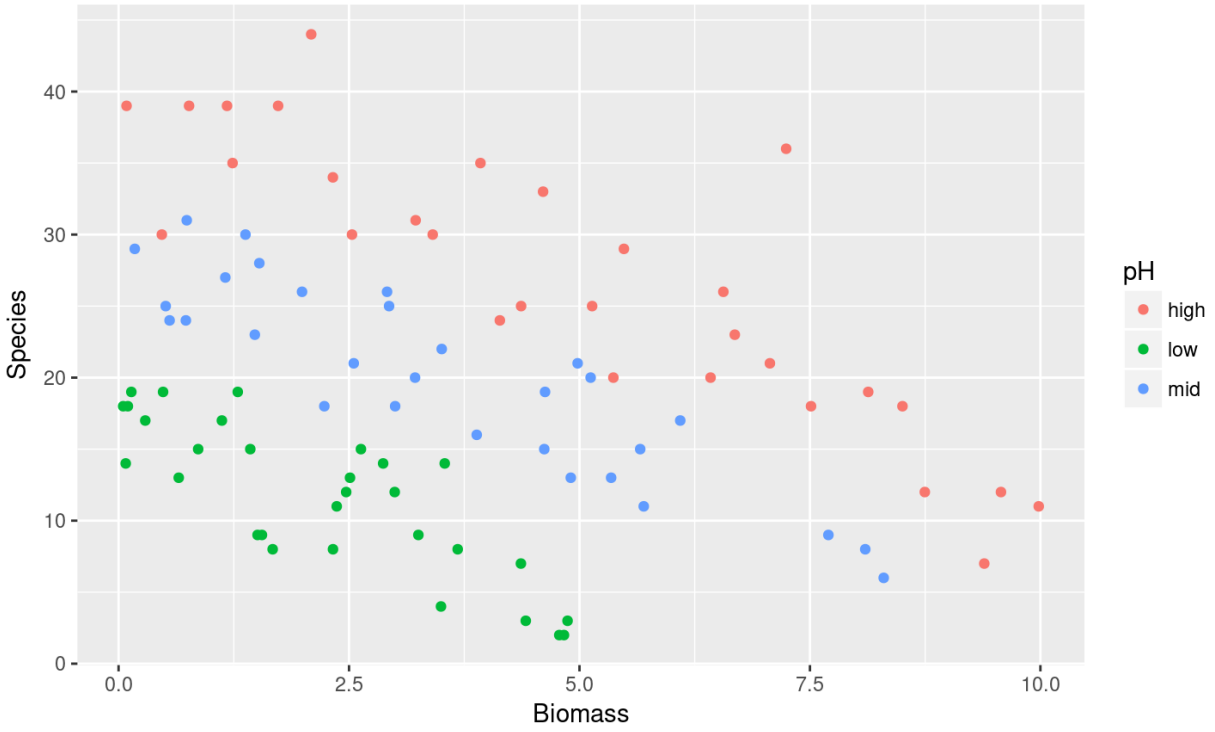
- faça o modelo cheio usando a familia de ligação **poisson(log)**
- avalie o sobre-dispersão do erro pela razão ''Residual deviance'' e ''degrees of freedom''
- se o valor da razão for muito maior que 1, ajuste o modelo cheio novamente com a família ''quasipoisson''
- compare os modelos simplificados com o mais complexo usando ''anova''
  - com ''poisson'' use o argumento ''test = "Chisq" ''
  - com ''quasipoisson'' use o argumento ''test = "F" ''
- retenha o modelo mínimo adequado
- retorne os coeficientes e preditos do modelo para escala original (antilog)

# Examplo: Contagem de espécies

- Biomassa e ph do solo estão relacionados à coexistência de espécies

| pH | Biomass | Species |
|---|---|---|
| **high** | 0.4692972 | 30 |
| **high** | 1.7308704 | 39 |
| **high** | 2.0897785 | 44 |
| **high** | 3.9257871 | 35 |
| **high** | 4.3667927 | 25 |
| **high** | 5.4819747 | 29 |

# Gráfico: riqueza de espécies




9/43

# Modelo

```
glm01 <- glm(Species ~ Biomass + pH + Biomass:pH, family = poisson, data= arv)
anova(glm01, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: Species
##
## Terms added sequentially (first to last)
##
##
##            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                          89     452.35
## Biomass     1   44.673        88     407.67 2.328e-11 ***
## pH          2  308.431        86      99.24 < 2.2e-16 ***
## Biomass:pH  2   16.040        84      83.20 0.0003288 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# *summary(glm01)*

```
##
## Call:
## glm(formula = Species ~ Biomass + pH + Biomass:pH, family = poisson,
##     data = arv)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4978  -0.7485  -0.0402   0.5575   3.2297
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.76812    0.06153  61.240  < 2e-16 ***
## Biomass      -0.10713    0.01249  -8.577  < 2e-16 ***
## pHlow        -0.81557    0.10284  -7.931 2.18e-15 ***
## pHmid        -0.33146    0.09217  -3.596 0.000323 ***
## Biomass:pHlow -0.15503    0.04003  -3.873 0.000108 ***
## Biomass:pHmid -0.03189    0.02308  -1.382 0.166954
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
```

# *summary(glm01)*

```
#(Dispersion parameter for poisson family taken to be 1)
#    Null deviance: 452.346  on 89  degrees of freedom
# Residual deviance:  83.201  on 84  degrees of freedom
# AIC: 514.39
```

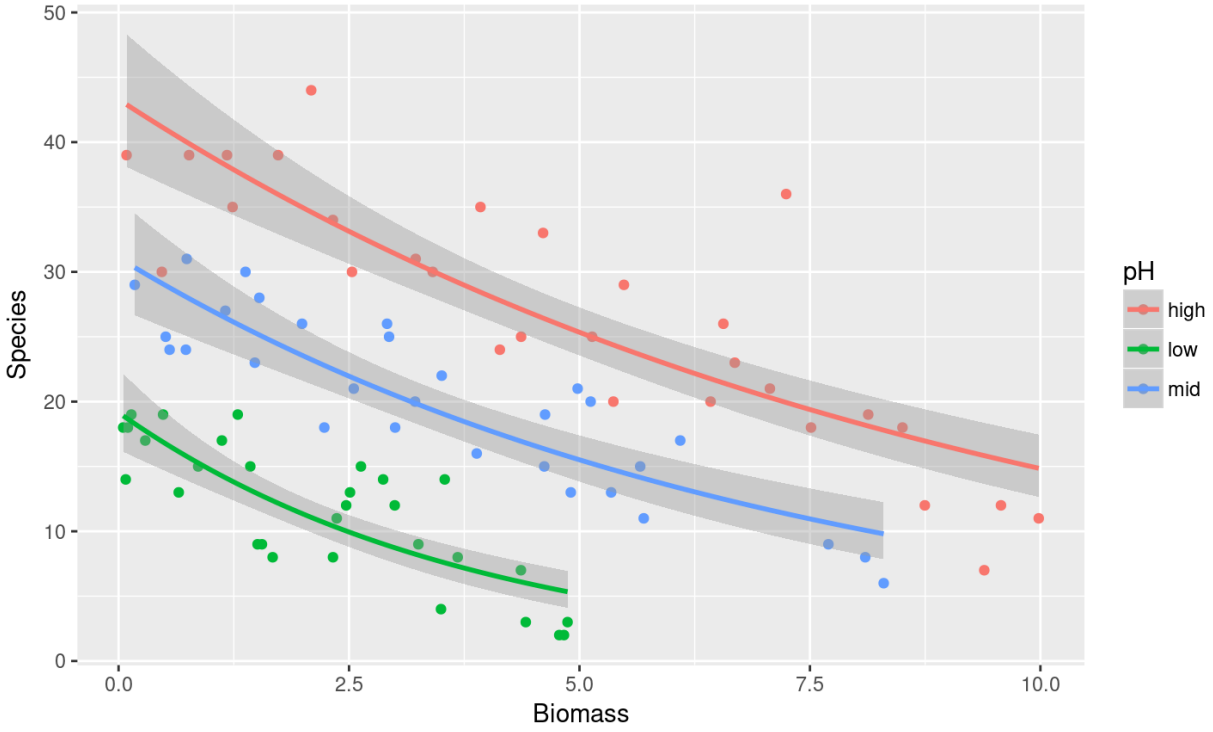
# Simplificando o Modelo

## ligação da família Poisson: Qui-quadrado

```
glm02 <- glm(Species ~ Biomass + pH, family = poisson, data= arv)
anova(glm01, glm02, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Species ~ Biomass + pH + Biomass:pH
## Model 2: Species ~ Biomass + pH
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        84     83.201
## 2        86     99.242 -2   -16.04 0.0003288 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Gráfico do nosso modelo



# Predito pelo modelo: preditor linear

- alto pH, biomassa 0.47 e 7.5

```
coefglm01 <- coef(glm01)
(bio01 <- coefglm01[[1]] + coefglm01[[2]] * 0.47)
```

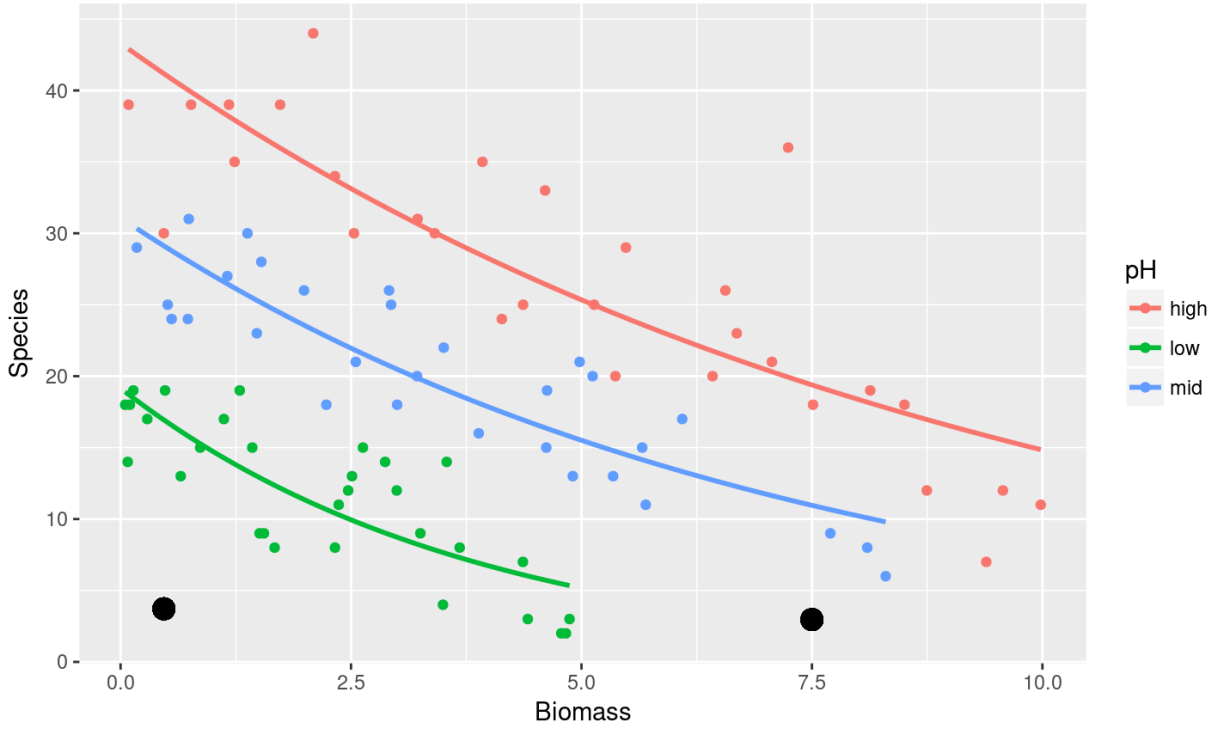
```
## [1] 3.717773
```

```
(bio02 <- coefglm01[[1]] + coefglm01[[2]] * 7.5)
```

```
## [1] 2.96465
```

# Predito pelo modelo: preditor linear

- alto pH, biomassa 0.47 e 7.5




16/43

# Predito pelo modelo: preditor linear

- alto pH, biomassa 0.47 e 7.5
- antilog: exp()

```
(exp(bio01))
```

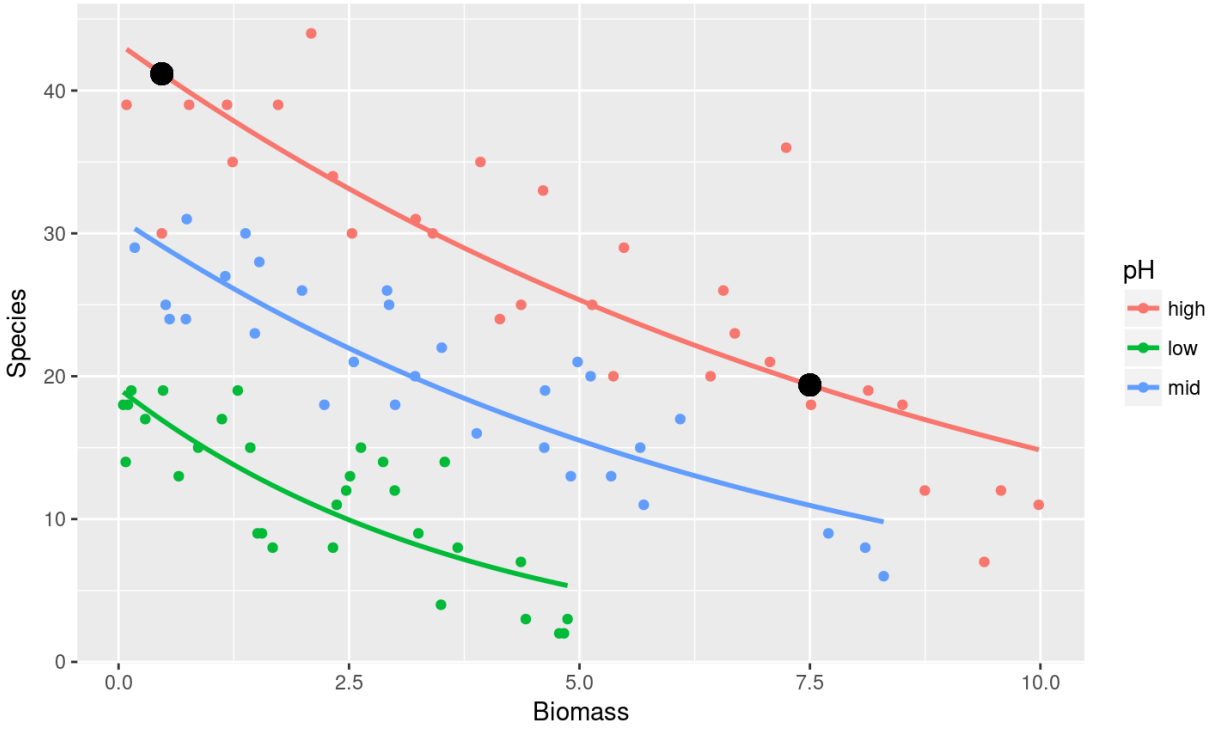
```
## [1] 41.17258
```

```
(exp(bio02))
```

```
## [1] 19.38792
```

# Predito pelo modelo: função inversa

$$exp(E_{(y)})$$

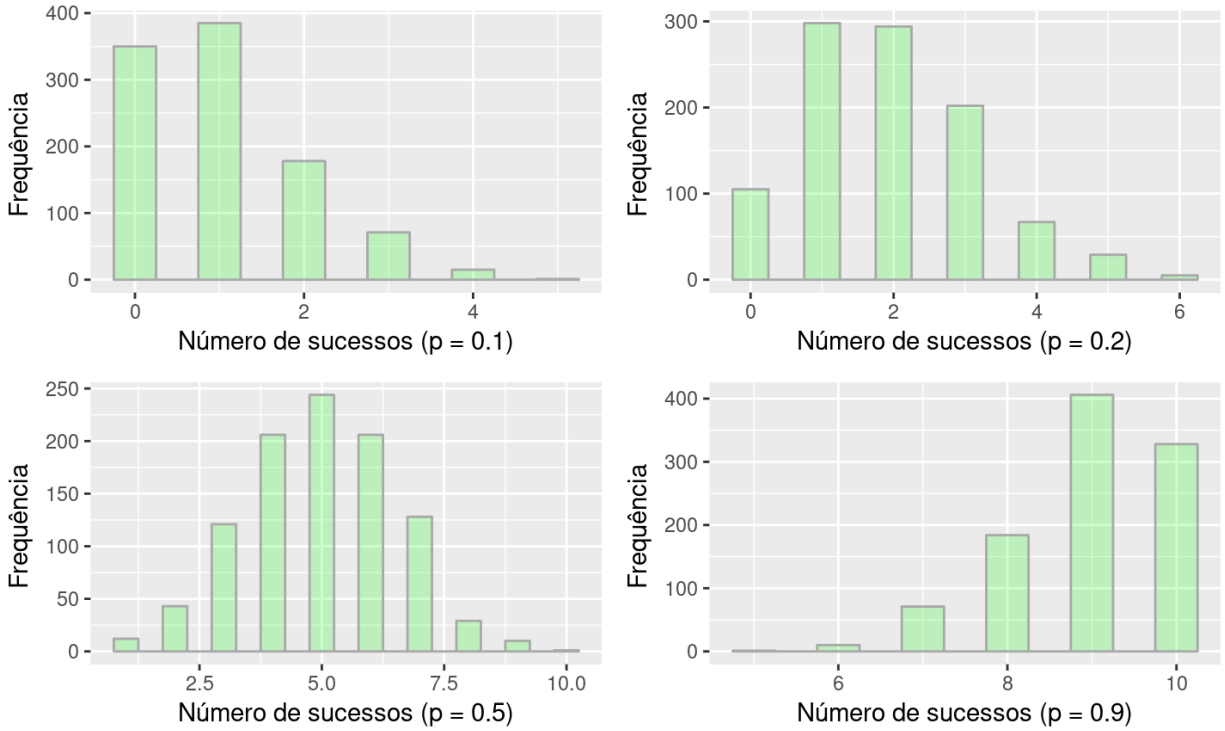


# Atividade

- Proporção (sucessos/tentativas)
- Binária:
  - sim x não
  - vivo x morto
  - germinou x não germinou

# GLM Binomial

# GLM Binomial: estrutura do erro



# GLM Binomial: logit

$$\eta = \log(\frac{a + bx}{1 - a + bx})$$

$$\eta = \frac{e^{a+bx}}{1 - e^{a+bx}}$$

$$x = +\infty; y_p = 1$$

$$x = -\infty; y_p = 0$$

# GLM Binomial: odds ratio (razão de chance)

- probabilidade : sucessos/tentativas
- chance: sucessos/falhas
  - sucessos/(tentativas - sucessos)

p = sucesssos

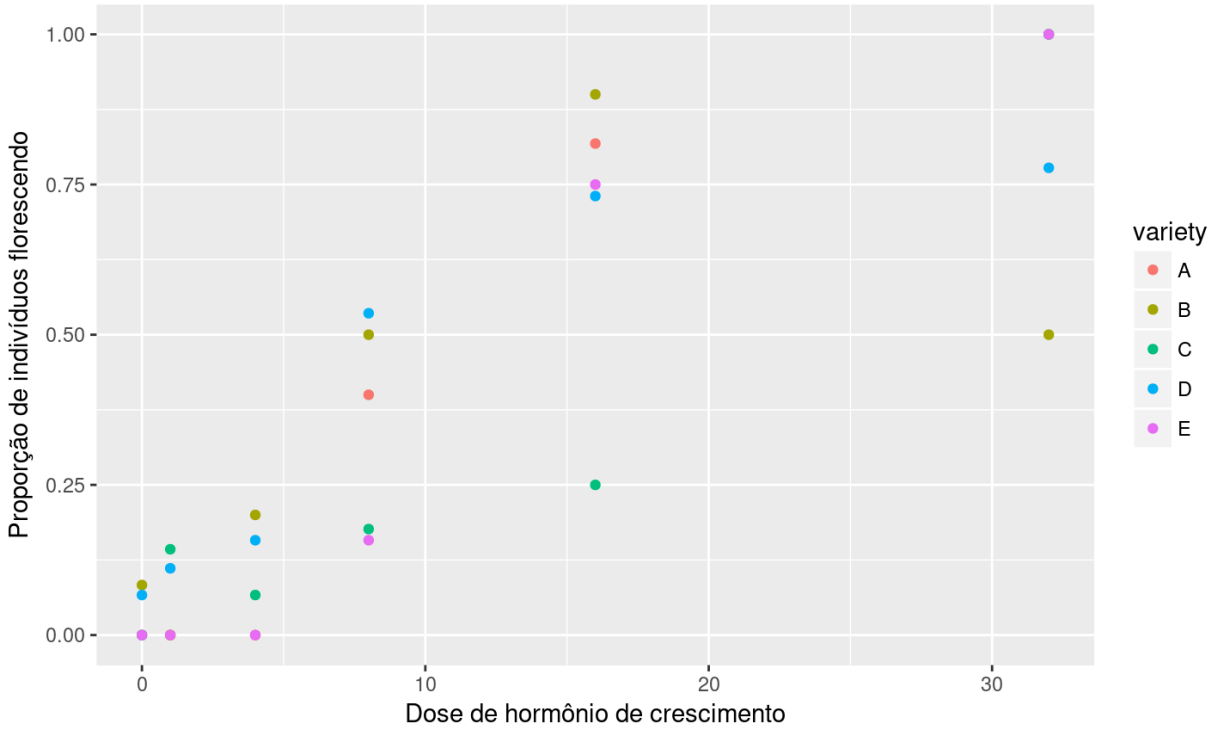
q = falha

$$log(\frac{p}{q}) = a + bx$$

# GLM binomial: florescer

| flowered | number | dose | variety | pf |
|---|---|---|---|---|
| **0** | 12 | 1 | A | 0.0000000 |
| **0** | 17 | 4 | A | 0.0000000 |
| **4** | 10 | 8 | A | 0.4000000 |
| **9** | 11 | 16 | A | 0.8181818 |
| **10** | 10 | 32 | A | 1.0000000 |
| **0** | 17 | 1 | B | 0.0000000 |

# Gráfico dos dados




25/43

# Variável preditora (sucesso, falhas)

```
(yb <- cbind(sucesso = flor$flowered, falha= flor$number-flor$flowered))
```

```
##       sucesso falha
##  [1,]       0    12
##  [2,]       0    17
##  [3,]       4     6
##  [4,]       9     2
##  [5,]      10     0
##  [6,]       0    17
##  [7,]       3    12
##  [8,]       6     6
##  [9,]       9     1
## [10,]       9     9
## [11,]       2    12
## [12,]       1    14
## [13,]       3    14
## [14,]       5    15
## [15,]      15     0
## [16,]       2    16
## [17,]       3    16
## [18,]      15    13
```

# GLM Binomial: ajustando o modelo

```
bin01 <- glm(yb ~ dose + variety + dose:variety, data=flor, family=binomial)
anova(bin01, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: yb
##
## Terms added sequentially (first to last)
##
##
##              Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                            29    303.350
## dose          1  197.098        28    106.252 < 2.2e-16 ***
## variety       4    9.483        24     96.769    0.0501 .
## dose:variety  4   45.686        20     51.083 2.863e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# GLM Binomial: ajustando o modelo

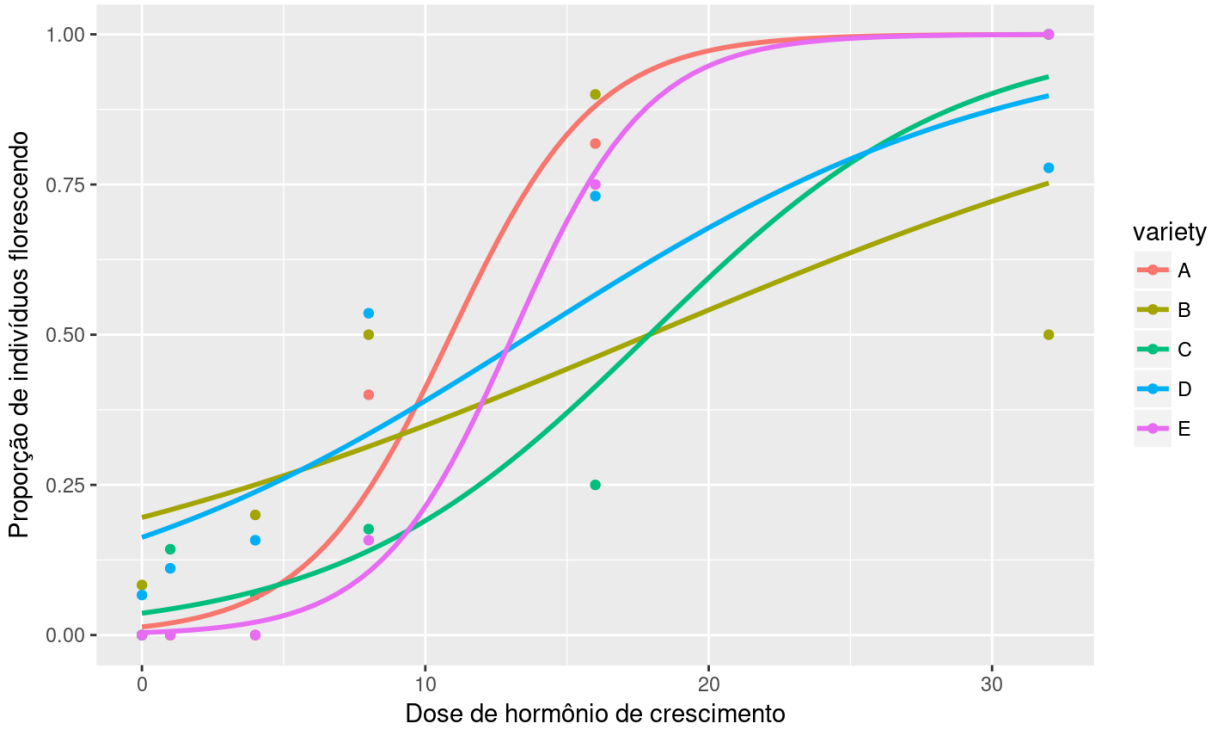
```
##
## Call:
## glm(formula = yb ~ dose + variety + dose:variety, family = binomial,
##     data = flor)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6648  -1.1200  -0.3769   0.5735   3.3299
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -4.59165    1.03215  -4.449 8.64e-06 ***
## dose           0.41262    0.10033   4.113 3.91e-05 ***
## varietyB       3.06197    1.09317   2.801 0.005094 **
## varietyC       1.23248    1.18812   1.037 0.299576
## varietyD       3.17506    1.07516   2.953 0.003146 **
## varietyE      -0.71466    1.54849  -0.462 0.644426
## dose:varietyB -0.34282    0.10239  -3.348 0.000813 ***
## dose:varietyC -0.23039    0.10698  -2.154 0.031274 *
## dose:varietyD -0.30481    0.10257  -2.972 0.002961 **
## dose:varietyE -0.00649    0.13292  -0.049 0.961057
```

# GLM Binomial: simplificando o modelo

```
bin02 <- glm(yb ~ dose + variety, data=flor, family=binomial)
anova(bin01, bin02, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: yb ~ dose + variety + dose:variety
## Model 2: yb ~ dose + variety
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        20     51.083
## 2        24     96.769 -4  -45.686 2.863e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# GLM Binomial: gráfico do modelo




30/43

# Predito pelo modelo: linear

```
predict(bin01)
```

```
##          1          2          3          4          5          6
## -4.1790352 -2.9411862 -1.2907208  2.0102098  8.6120711 -1.4598882
##          7          8          9         10         11         12
## -1.2504982 -0.9713115 -0.4129381  0.7038087 -3.1769379 -2.6302485
##         13         14         15         16         17         18
## -1.9013292 -0.4434907  2.4721863 -1.3087930 -0.9853868 -0.5541786
##         19         20         21         22         23         24
##  0.3082377  2.0330705 -4.9001833 -3.6818046 -2.0572996  1.1917103
##         25         26         27         28         29         30
##  7.6897302 -4.5916515 -1.5296848 -3.3591677 -1.4165950 -5.3063095
```

# Predito pelo modelo: antilogit

```
1/(1+ 1/exp(predict(bin01)))
```

```
##           1           2           3           4           5           6
## 0.015082316 0.050154735 0.215730827 0.881864884 0.999818136 0.188484430
##           7           8           9          10          11          12
## 0.222613918 0.274619176 0.398207832 0.669031659 0.040042874 0.067216871
##          13          14          15          16          17          18
## 0.129958109 0.390909521 0.922168829 0.212688893 0.271824227 0.364895476
##          19          20          21          22          23          24
## 0.576455053 0.884225776 0.007390197 0.024559161 0.113316872 0.767046813
##          25          26          27          28          29          30
## 0.999542708 0.010034395 0.178039802 0.033596235 0.195195932 0.004935716
```

# GLM: ajuste de modelo proporção

- ajuste a variável resposta (sucesso, falha)
- use a familia de ligação **binomial(logit)**
- avalie o sobre-dispersão do erro pela razão ''Residual deviance'' e ''degrees of freedom''
- razão > 1, use família ''quasibinom''
- busque o modelo mínimo adequado com ''anova''
  - ''poisson'' use ''test = "Chisq" ''
  - ''quasipoisson'' use ''test = "F" ''
- retenha o modelo mínimo adequado
- retorne os coeficientes e preditos do modelo para escala original (antilogit)

# Atividade: GLM Binomial

# Modelos Lineares Mistos

coeficientes da regressão são variáveis aleatórias

- modelo mais simples: intercepto é uma variável aleatória

$$y = \alpha + \beta x + \epsilon$$

$$\alpha = N(\hat{\alpha}, \sigma^2)$$

- estima-se outra variância, associada à variável randômica
- $\alpha$ não é um valor e sim um distribuição de probabilidades de valores

# LMM: riqueza na praia

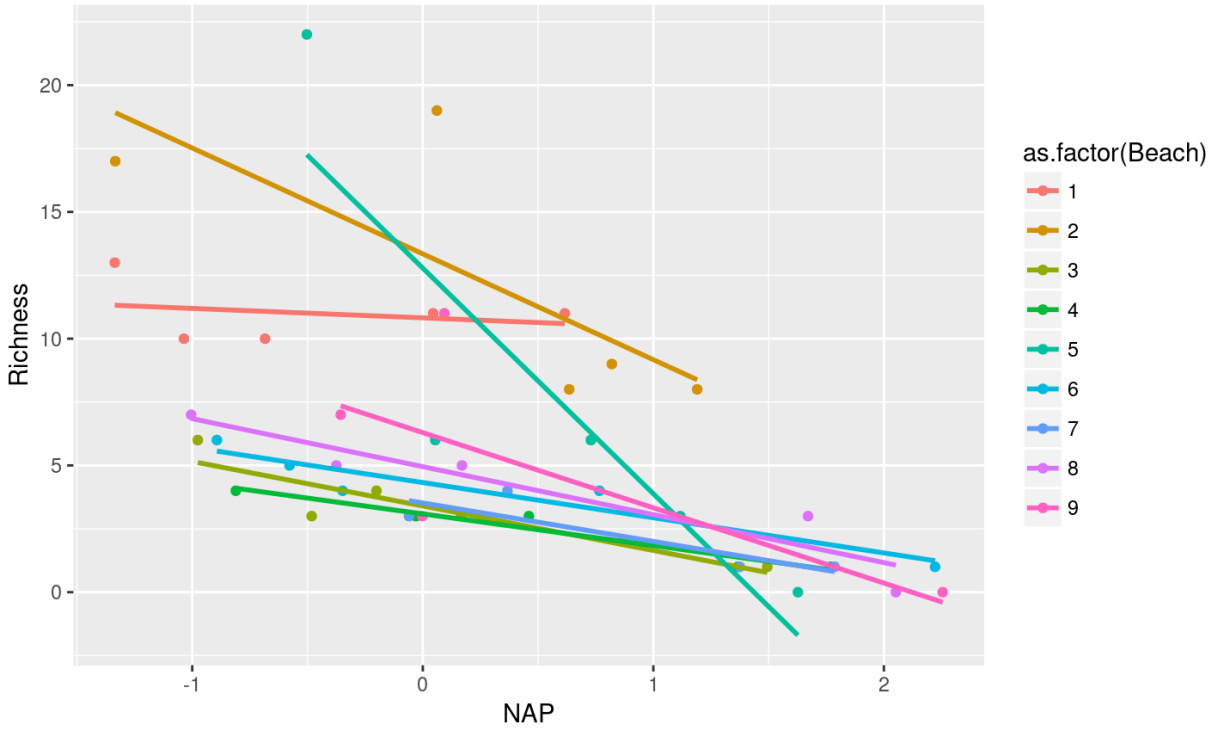
A riqueza da macrofauna varia em função da altura relativa ao nível médio da mare

| Sample | Richness | Exposure | NAP | Beach |
|---|---|---|---|---|
| **1** | 11 | 10 | 0.045 | 1 |
| **2** | 10 | 10 | -1.036 | 1 |
| **3** | 13 | 10 | -1.336 | 1 |
| **4** | 11 | 10 | 0.616 | 1 |
| **5** | 10 | 10 | -0.684 | 1 |
| **6** | 8 | 8 | 1.190 | 2 |

**## LMM: r** iqueza na p raia

# LMM: riqueza na praia

Uma possibilidade: 9 regressões



# LMM: riqueza na praia

há um efeito aleatório da praia

- as observações dentro de cada praia não são independentes

```
library(lme4)
```

```
## Loading required package: Matrix
```

```
lmm01 <- lmer(Richness ~ NAP + (1|Beach), data=praia)
```

# LMM: riqueza da praia

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Richness ~ NAP + (1 | Beach)
##    Data: praia
##
## REML criterion at convergence: 239.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.4227 -0.4848 -0.1576  0.2519  3.9794
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Beach    (Intercept) 8.668    2.944
##  Residual             9.362    3.060
## Number of obs: 45, groups:  Beach, 9
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   6.5819     1.0958   6.007
## NAP          -2.5684     0.4947  -5.192
##
```

# LMM: modelo mínimo adequado

## Comparando modelos mistos

- os modelos devem ser comparados por "ML"
- devem ser apresentados com "RML"

```
lmm01r <- lmer(Richness ~ NAP + (1|Beach), data=praia, REML =  FALSE)
lmm00r <- lmer(Richness ~ 1 + (1|Beach), data=praia, REML = FALSE)
anova(lmm00r, lmm01r)
```

```
## Data: praia
## Models:
## lmm00r: Richness ~ 1 + (1 | Beach)
## lmm01r: Richness ~ NAP + (1 | Beach)
##        Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## lmm00r  3 269.30 274.72 -131.65   263.30
## lmm01r  4 249.83 257.06 -120.92   241.83 21.474      1  3.586e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

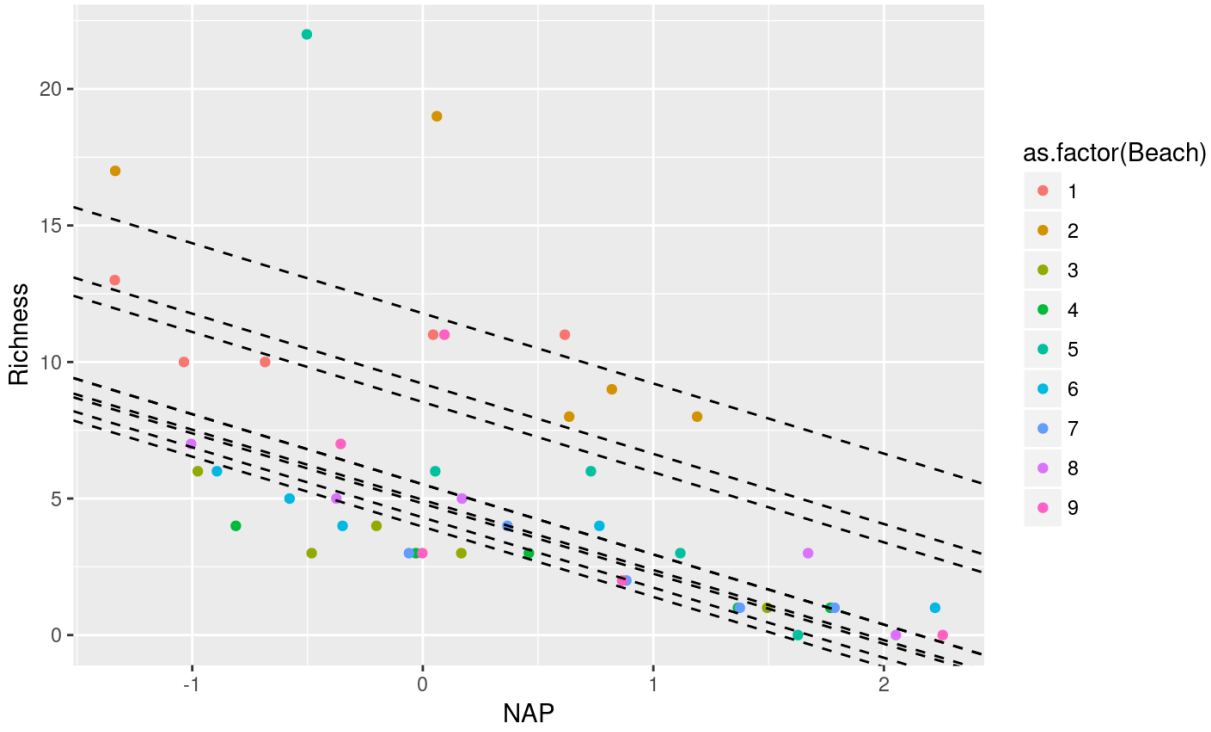
40/43

# LMM: resultado do modelo

```
(coeflmm <- coef(lmm01))
```

```
## $Beach
##   (Intercept)     NAP
## 1    9.203412 -2.5684
## 2   11.781501 -2.5684
## 3    3.966113 -2.5684
## 4    4.306275 -2.5684
## 5    8.532072 -2.5684
## 6    4.952491 -2.5684
## 7    4.816416 -2.5684
## 8    5.520228 -2.5684
## 9    6.158529 -2.5684
##
## attr(,"class")
## [1] "coef.mer"
```

# LMM: resultado do modelo




42/43

# Atividade: LMM intercepto e inclinação