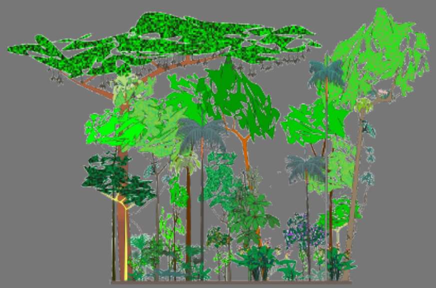
Modelos Lineares Generalizados

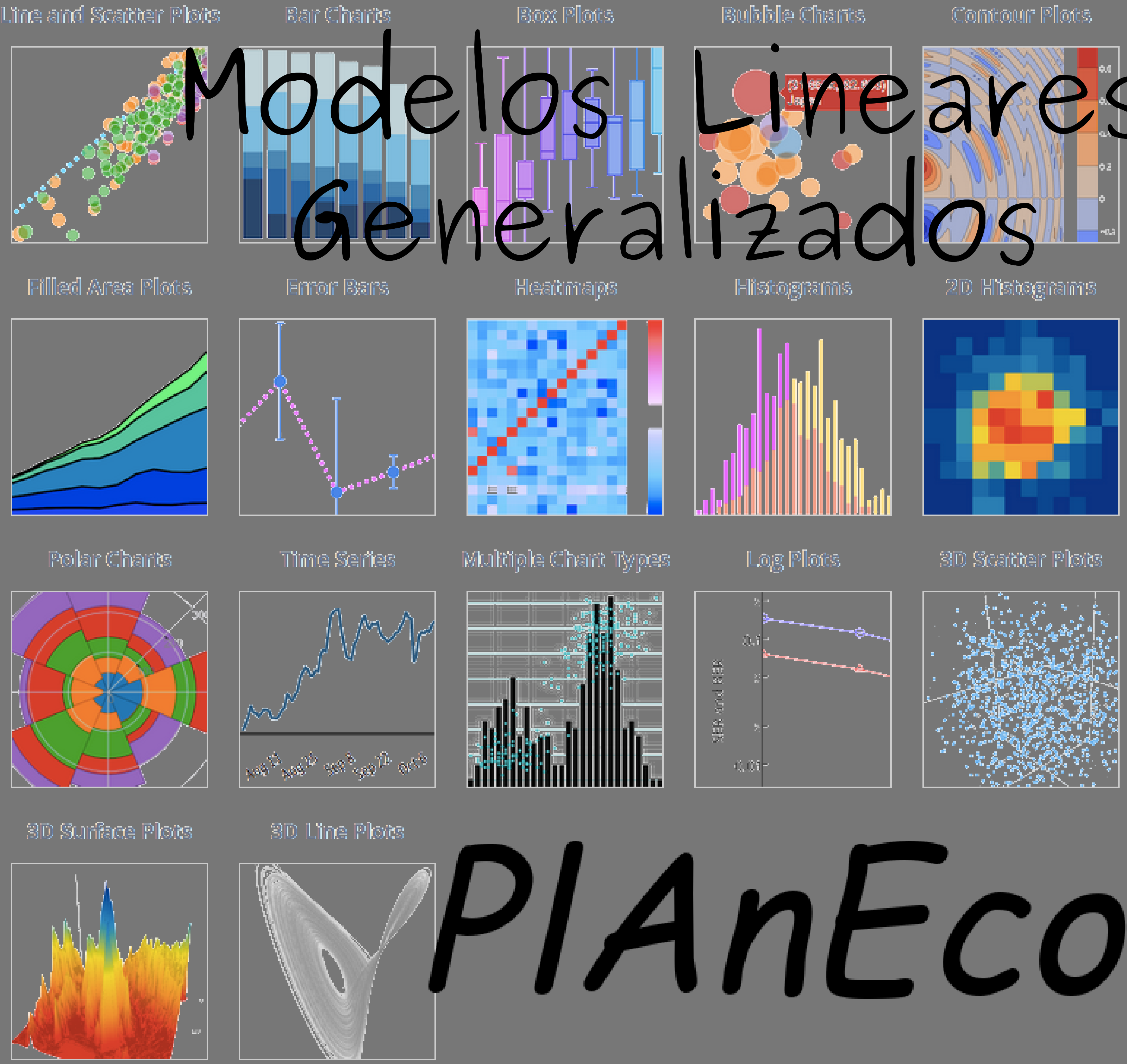
unificação metodológica



Alexandre Adalardo de Oliveira

PlanECO 2019

Modelos Lineares Generalizados



PlAnEco

Modelos Lineares Generalizados (GLMs)

Conceitos Tratados

- estrutura do erro
- preditora linear
- função de ligação
- inverso da função de ligação

Modelos Lineares Generalizados (GLMs)

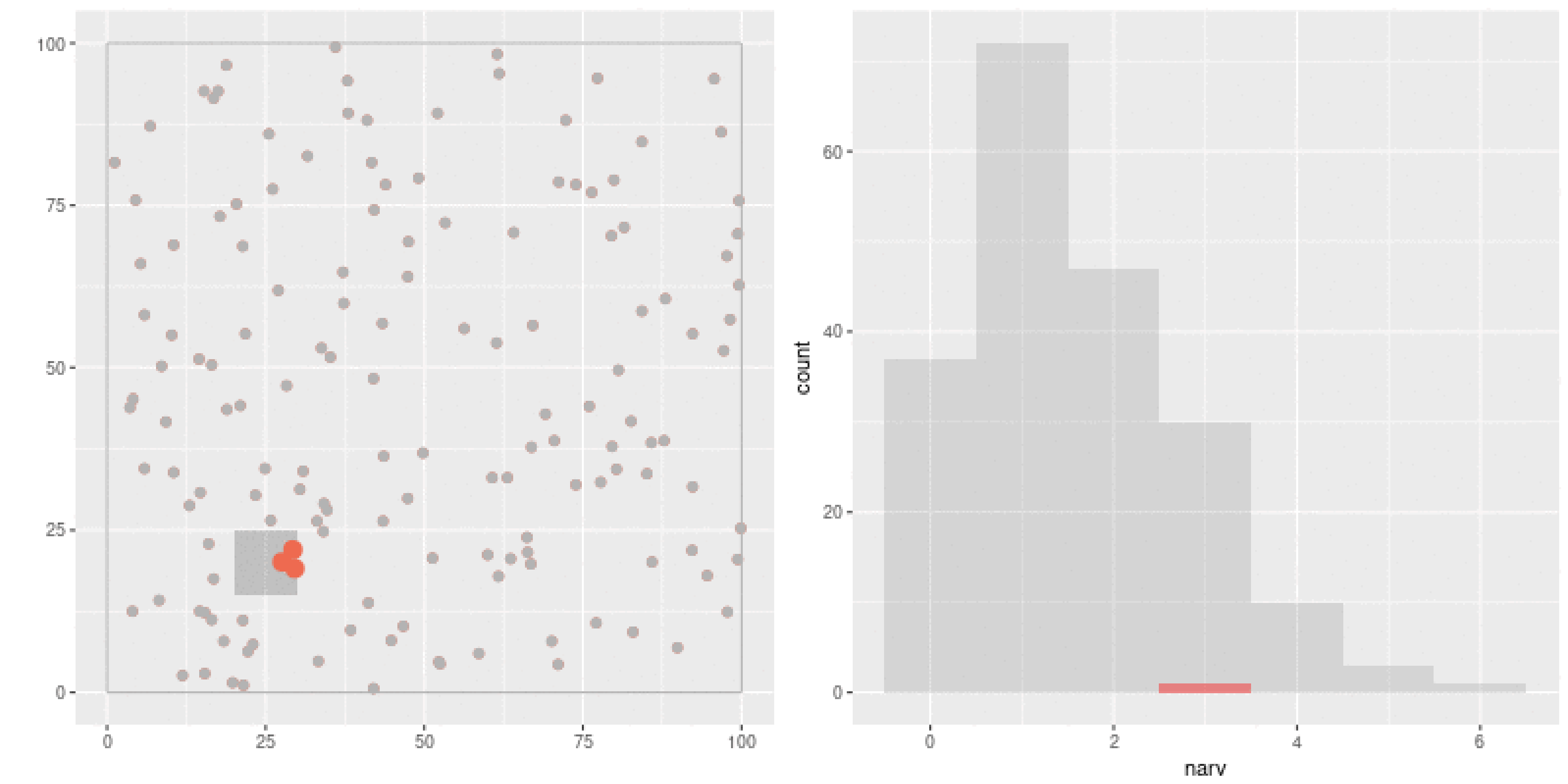
Outros Conceitos

- Verossimilhança (MLE)
- Quasi-Likelihood
- deviance

Estrutura do Erro

Contagem de árvores em Parcelas

Completa Aleatoriedade Espacial



GLMs: conceitos

Função de ligação

- a função de ligação é aplicada na $\boldsymbol{Esperança}$ (valor esperado) da variável resposta aleatória:

$$\eta = g(\mu)$$

Preditor Linear

- o preditor η é a resposta linear da soma de uma ou mais preditoras

$$\eta = \alpha + \sum_{j=1}^{p} \beta_j x_{ij}$$

Função inversa

- 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

LMs são GLMs com função de ligação identidade

$$\eta = I(\hat{Y})$$

$$\eta = \hat{Y}$$

GLMs x Tranformar LMs

Modelo Transformado de LMs

$$f(y)$$
$$f(\hat{Y} + \epsilon)$$

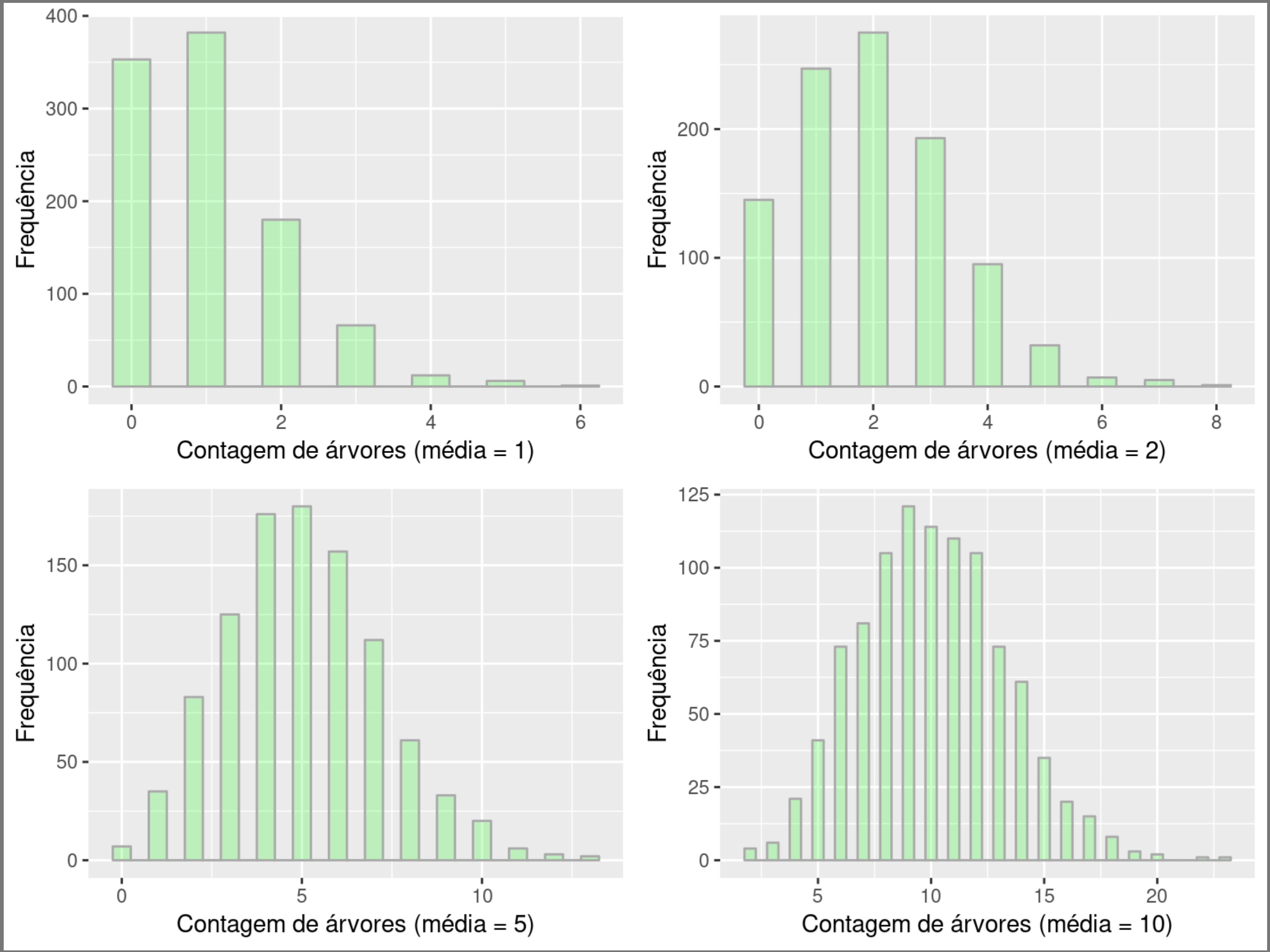
- transforma o valor observado (predito mais o resíduo)

Função de Ligação

$$f(\hat{Y})$$

- tranforma apenas a experança da variável aleatória
- preserva a estrutura do resíduo
- modela a estrutura do resíduo na escala original (Poisson, Binomial...)

Estruturas de erro



Dados de Contagem

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

Função de ligação

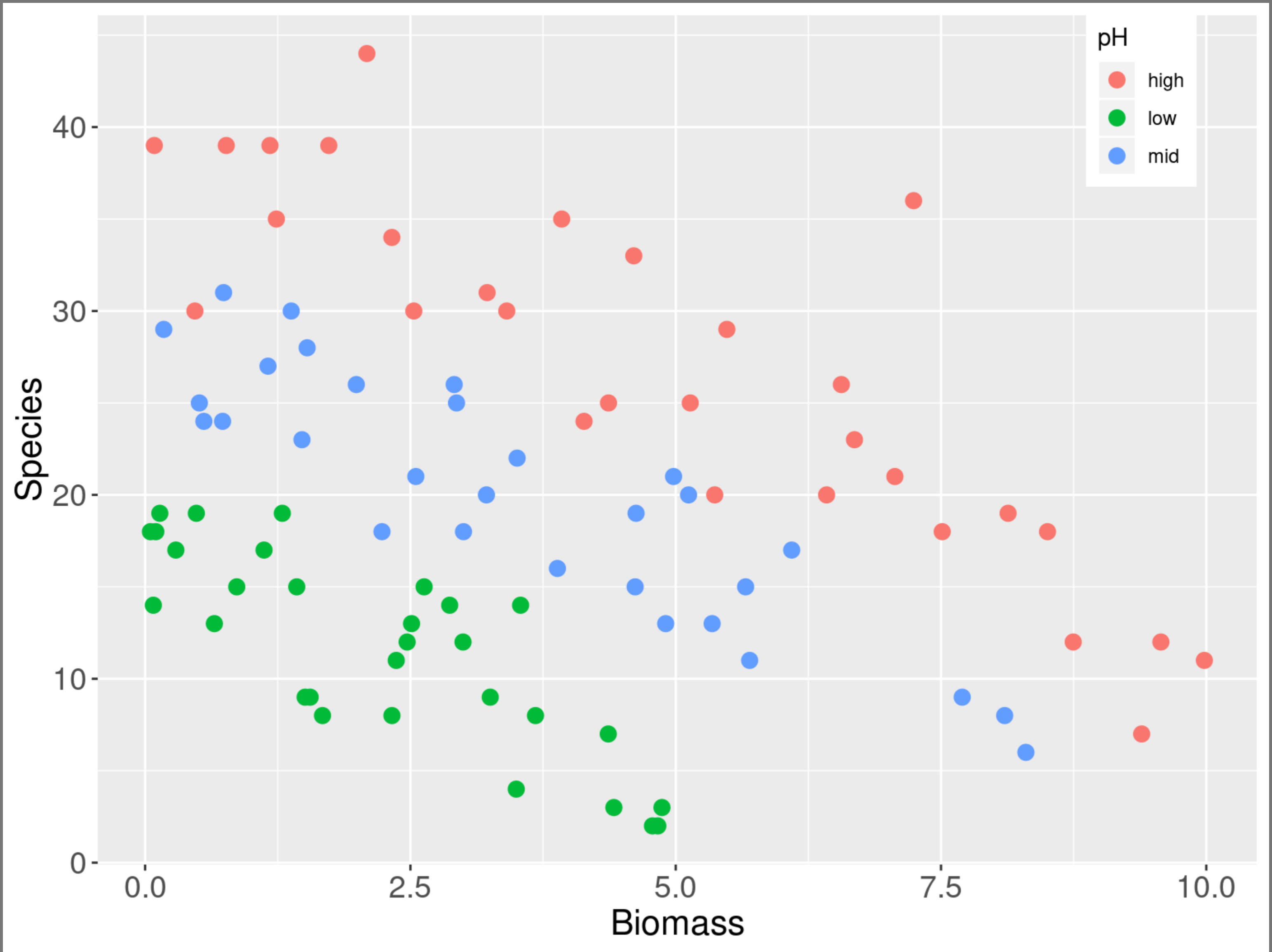
$$\log(\alpha + \sum \beta_i x_i)$$

Examplo: Contagem de espécies

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

	pH	Biomass	Species
1	high	0.4692972	30
2	high	1.7308704	39
3	high	2.0897785	44
31	mid	0.1757627	29
32	mid	1.3767783	30
33	mid	2.5510426	21
61	low	0.1008479	18
62	low	0.1385961	19
63	low	0.8635151	15

Gráfico: riqueza de espécies



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 '
```

summary(glm01)

```
##
## Call:
## glm(formula = Species ~ Biomass + pH + Biomass:pH, family =
##     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 '
##
## (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
##
## Number of Fisher Scoring iterations: 4
```

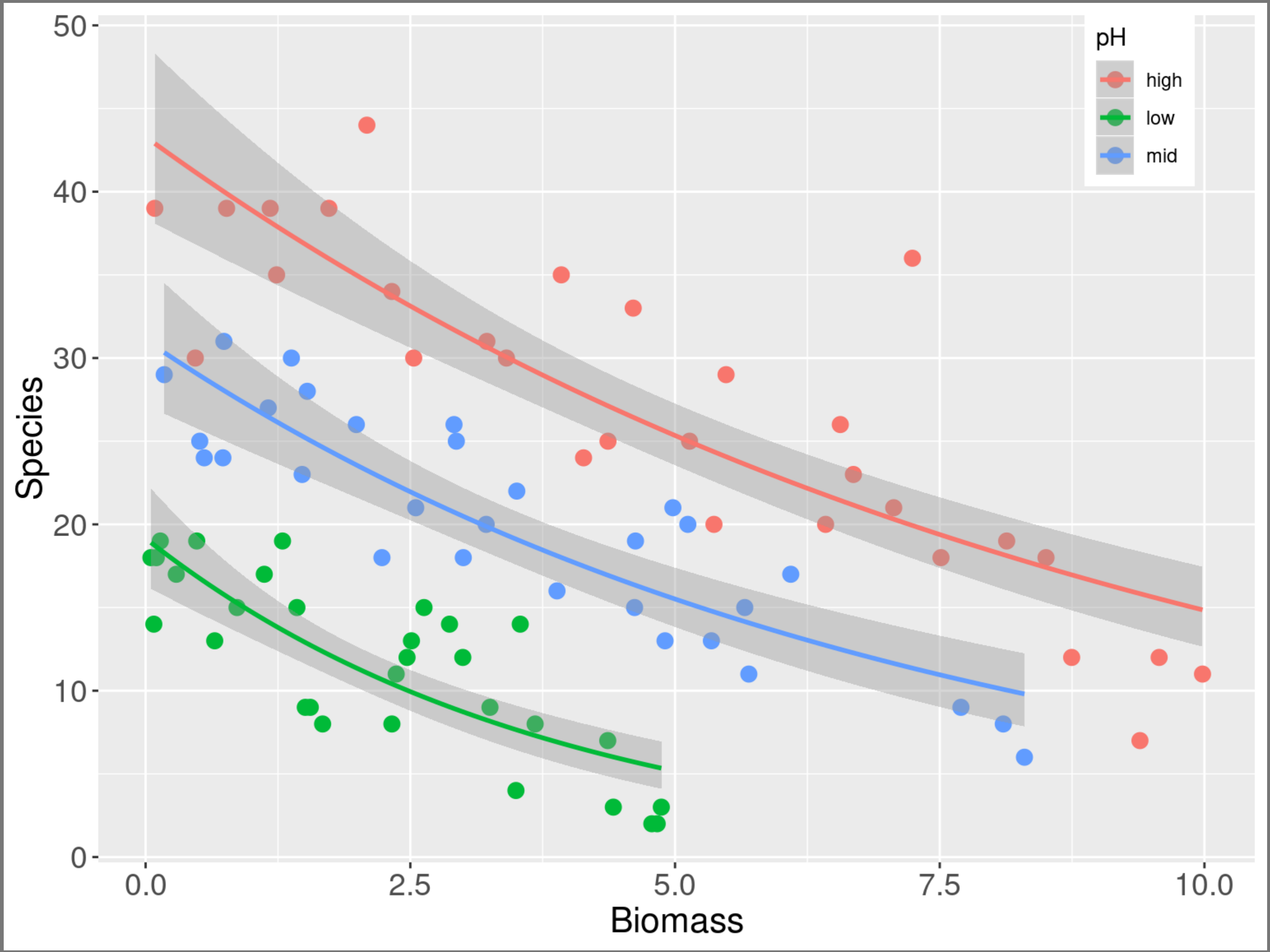
Simplificando o Modelo

Modelos 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

- alto pH, biomassa 0.47 e 0.75

```
## (Intercept)     Biomass
##   3.7681236  -0.1071298
```

$$3.77 - 0.107 * 0.47$$

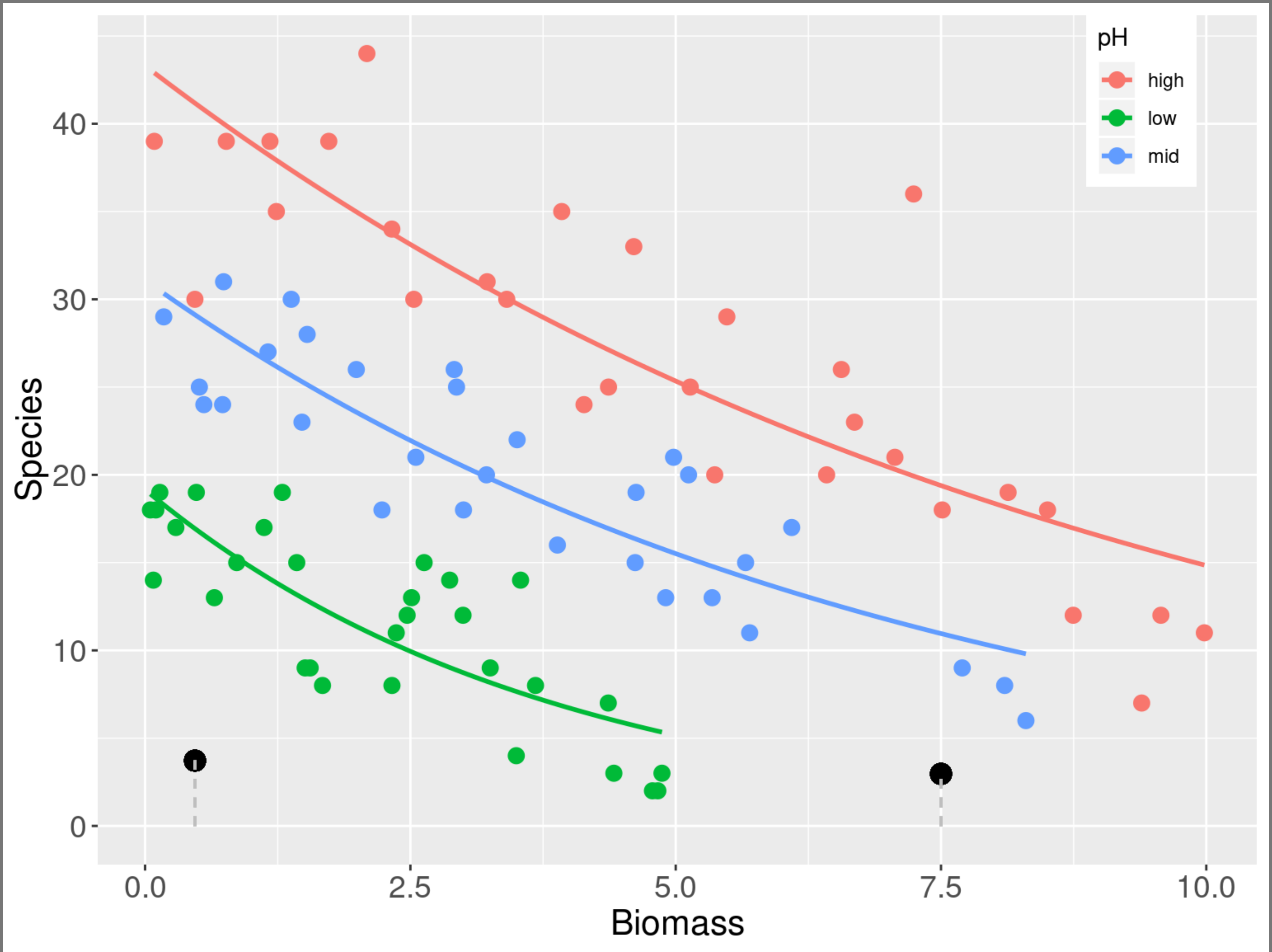
```
## [1] 3.717773
```

$$3.77 - 0.107 * 0.75$$

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

Predito pelo modelo

- alto pH, biomassa 0.47 e 7.5



Predito pelo modelo: preditor linear

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

$$exp(3.717)$$

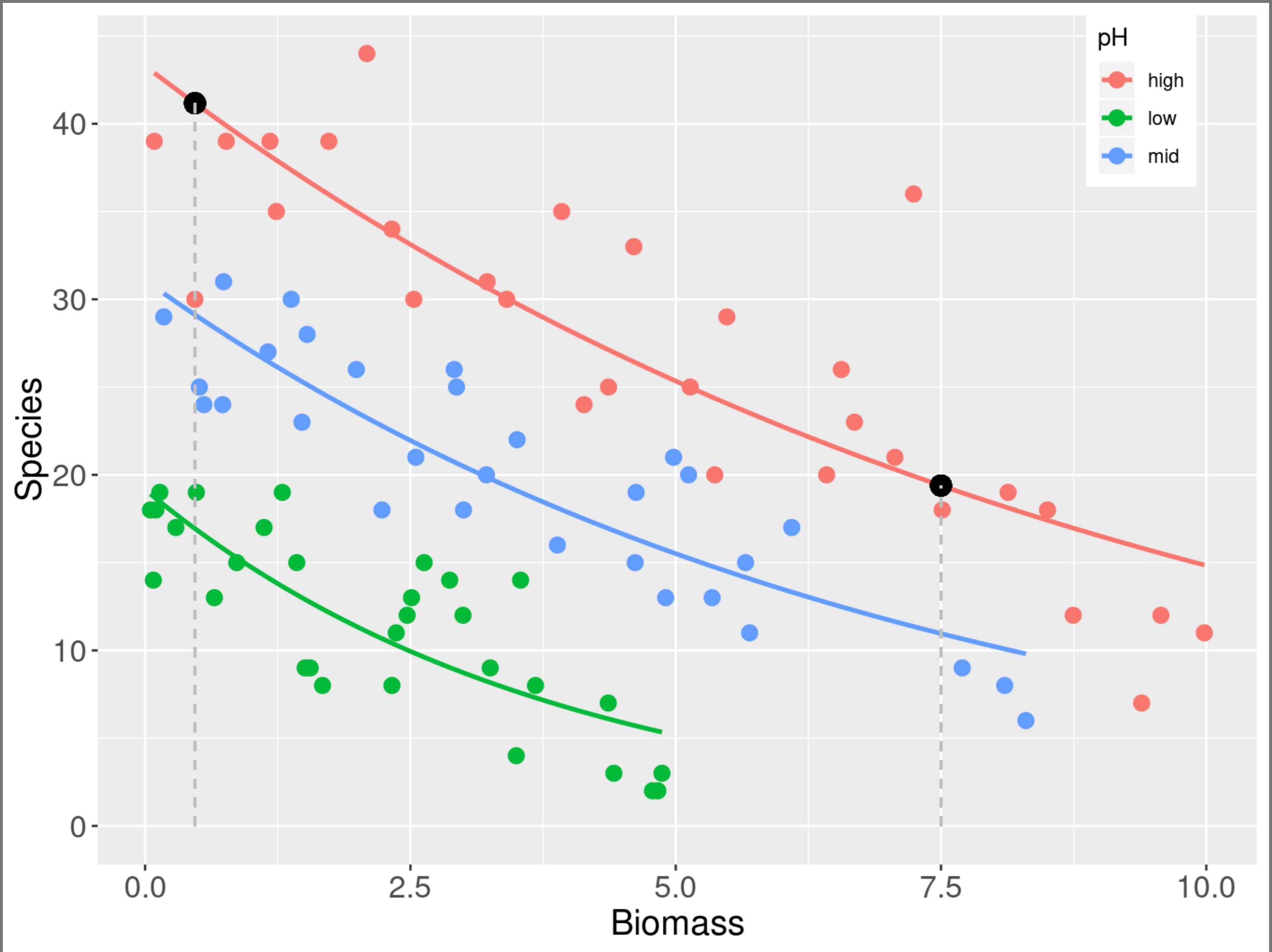
```
## [1] 41.17258
```

$$exp(2.964)$$

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

Predito: função inversa

$$exp(\eta) \;\; ou \;\; exp(\hat{\alpha} + \hat{\beta} * x)$$



pH
high
low
mid
Species
Biomass
0
10
20
30
40
0.0
2.5
5.0
7.5
10.0

GLM: modelo de contagem

- faça o modelo cheio usando a familia de ligação poisson(log)
- avalie o sobre-dispersão do erro pela razão: Residual deviance/degrees of freedom
- se o valor da razão for muito maior que 1, ajuste o modelo cheio novamente com a família quasipoisson
- se a sobredispersão persistir uma alternativa e modelar o resíduo com a binomial negativa (um parâmetro a mais relacionado à agregação)
- 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)

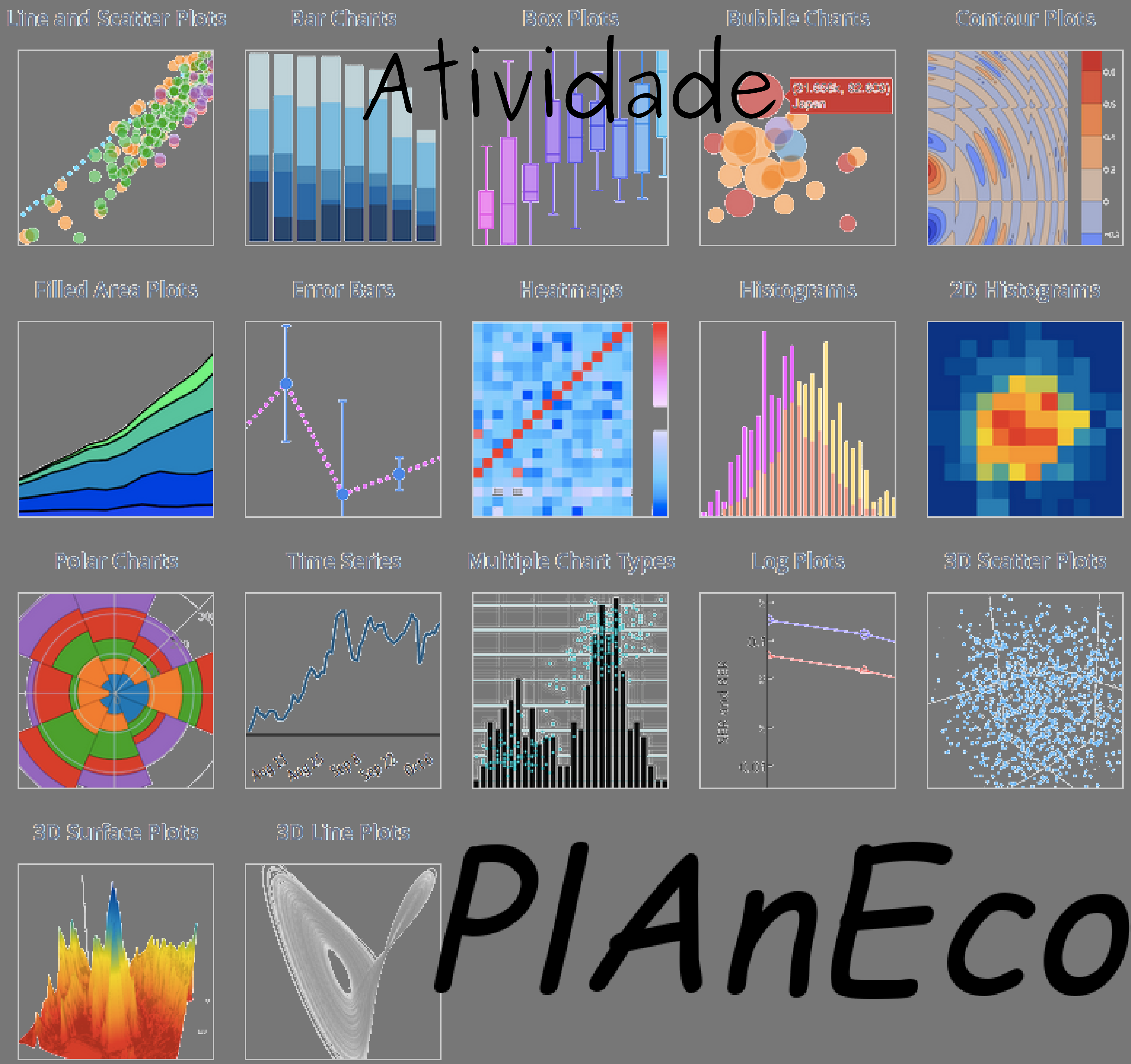
summary(glm01)

```
#(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
```

O que fazer quando o parâmetro de dispersão é diferente de 1?

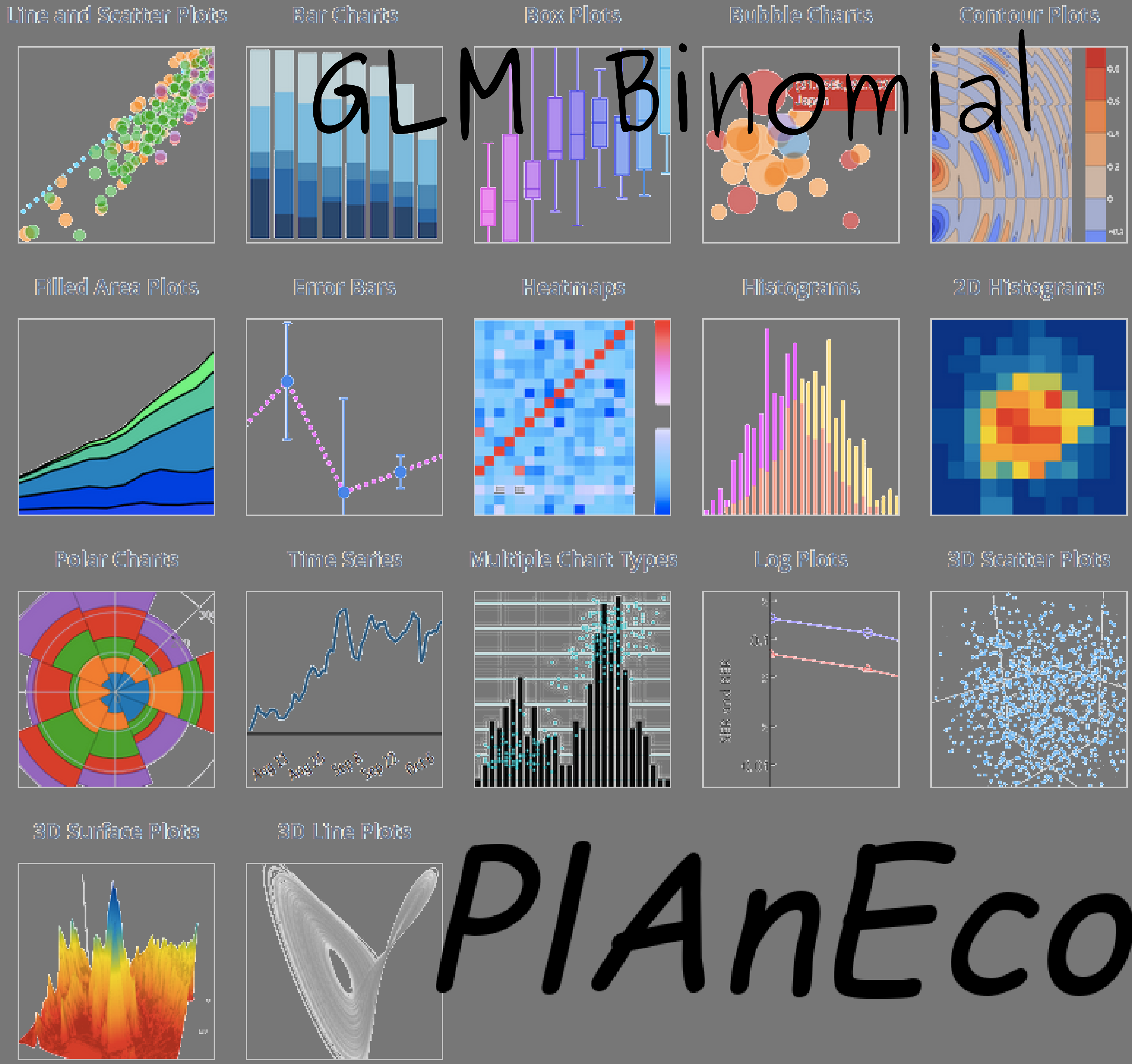
- usar `family= quasipoisson`
- usar modelo com distribuição binomial negativa : `glm.nb()` pacote `MASS`

Atividade



PlAnEco

GLM Binomial



PlAnEco

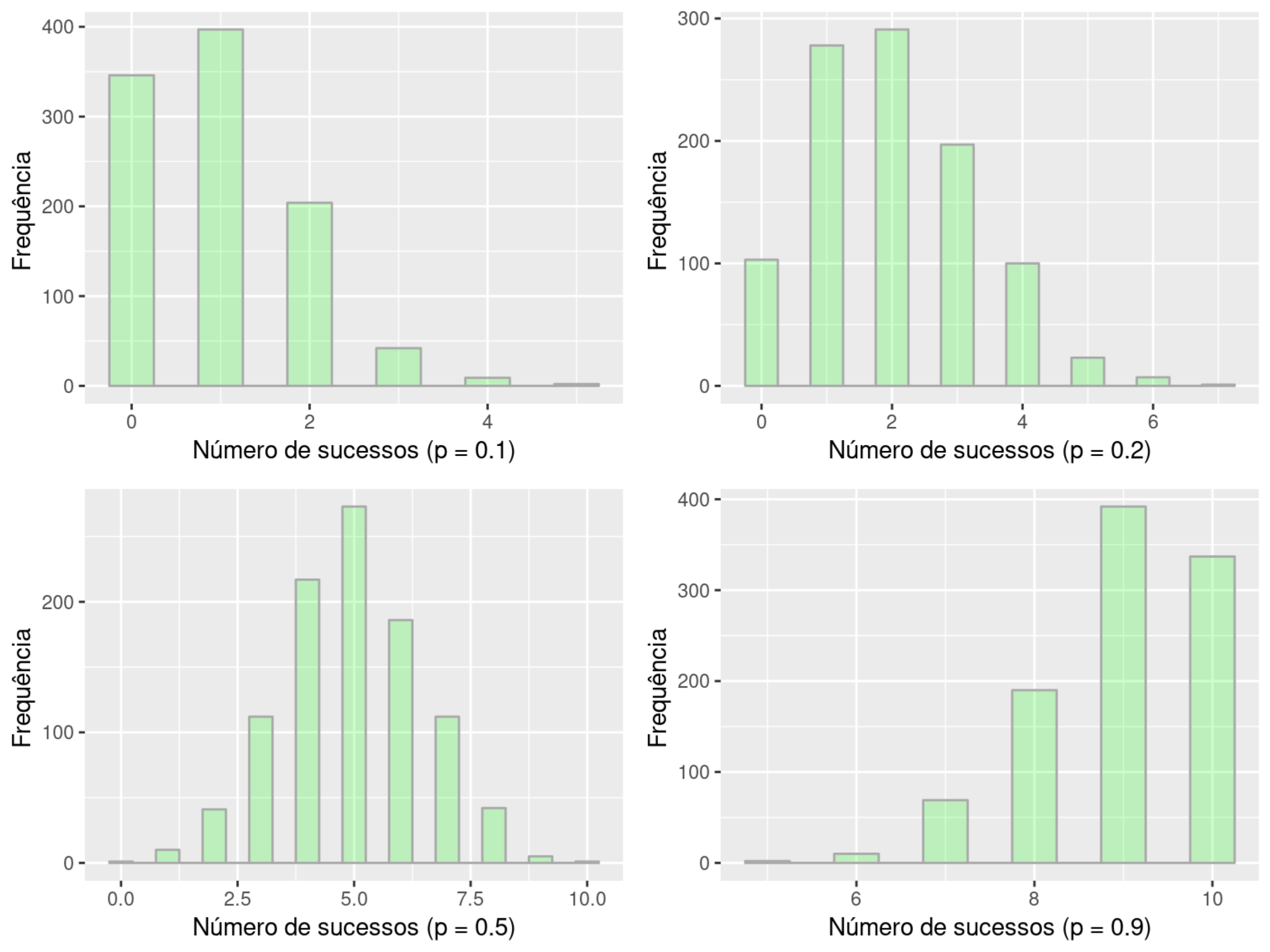
GLM Binomial

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

limite [0–1]

variância depende da média

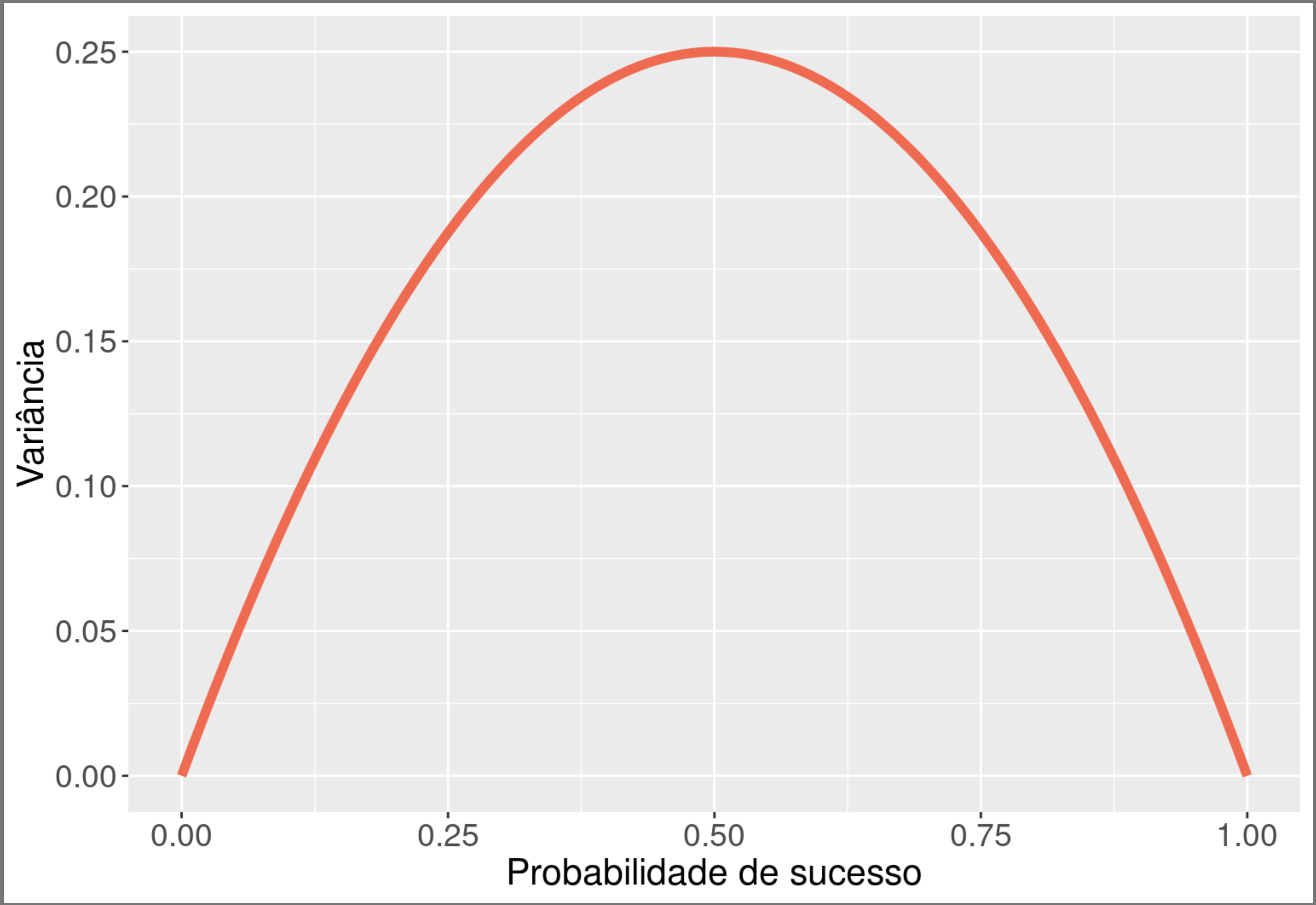
GLM Binomial: estrutura do erro



Frequência
Número de sucessos (p = 0.1)
Número de sucessos (p = 0.2)
Número de sucessos (p = 0.5)
Número de sucessos (p = 0.9)

GLM Binomial: variância

$$\sigma^2 = npq$$



GLM Binomial: odds ratio

Razão de Chance

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

n = tentativas

p = sucesssos

q = falha

q = n − p

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

GLM Binomial: logit

Log da Razão de Chance (log(Odds Ratio))

$$\eta = \log(\frac{p}{q})$$

$$\eta = \log(\frac{p}{1-p})$$

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

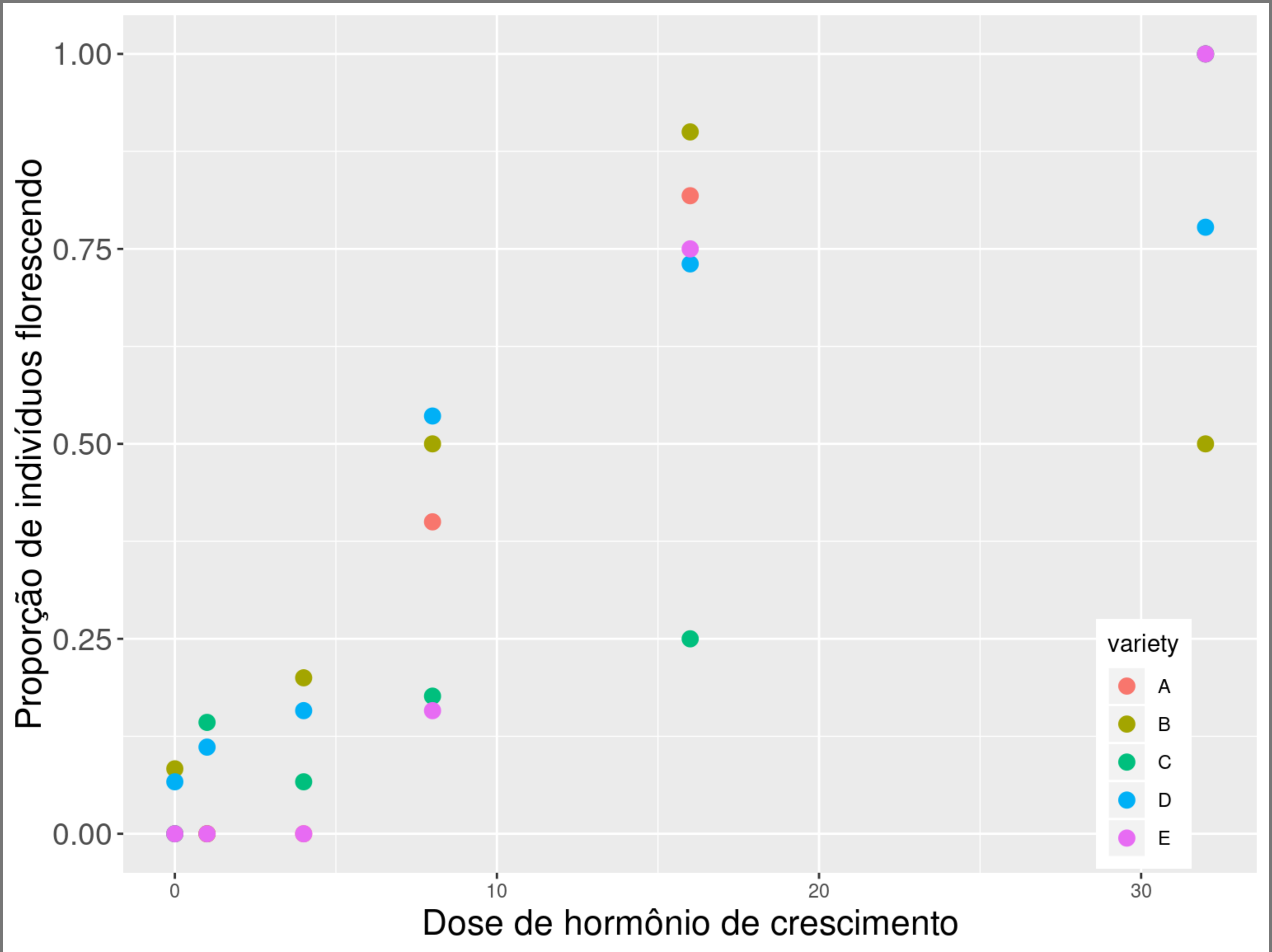
$$x \to \infty; y_p \to 1$$

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

GLM binomial: florescer

flowered	number	dose	variety
0	12	1	A
0	17	4	A
0	17	1	B
3	15	4	B
2	14	1	C
1	15	4	C
2	18	1	D
3	19	4	D

Gráfico dos dados



Variável resposta

- combinação de sucesso e falhas

sucesso	falha
0	12
0	17
4	6
9	2
10	0
0	17
3	12
6	6
9	1
9	9
2	12
1	14
3	14
5	15
15	0
2	16

sucesso	falha
3	16
15	13
19	7
21	6
0	13
0	15
3	16
15	5
17	0
0	11
1	11
0	17
1	14
0	10

GLM Binomial: ajustando o modelo

(sucesso, falha) ~ dose + variety + dose:variety

```
## 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 = bino
##     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
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

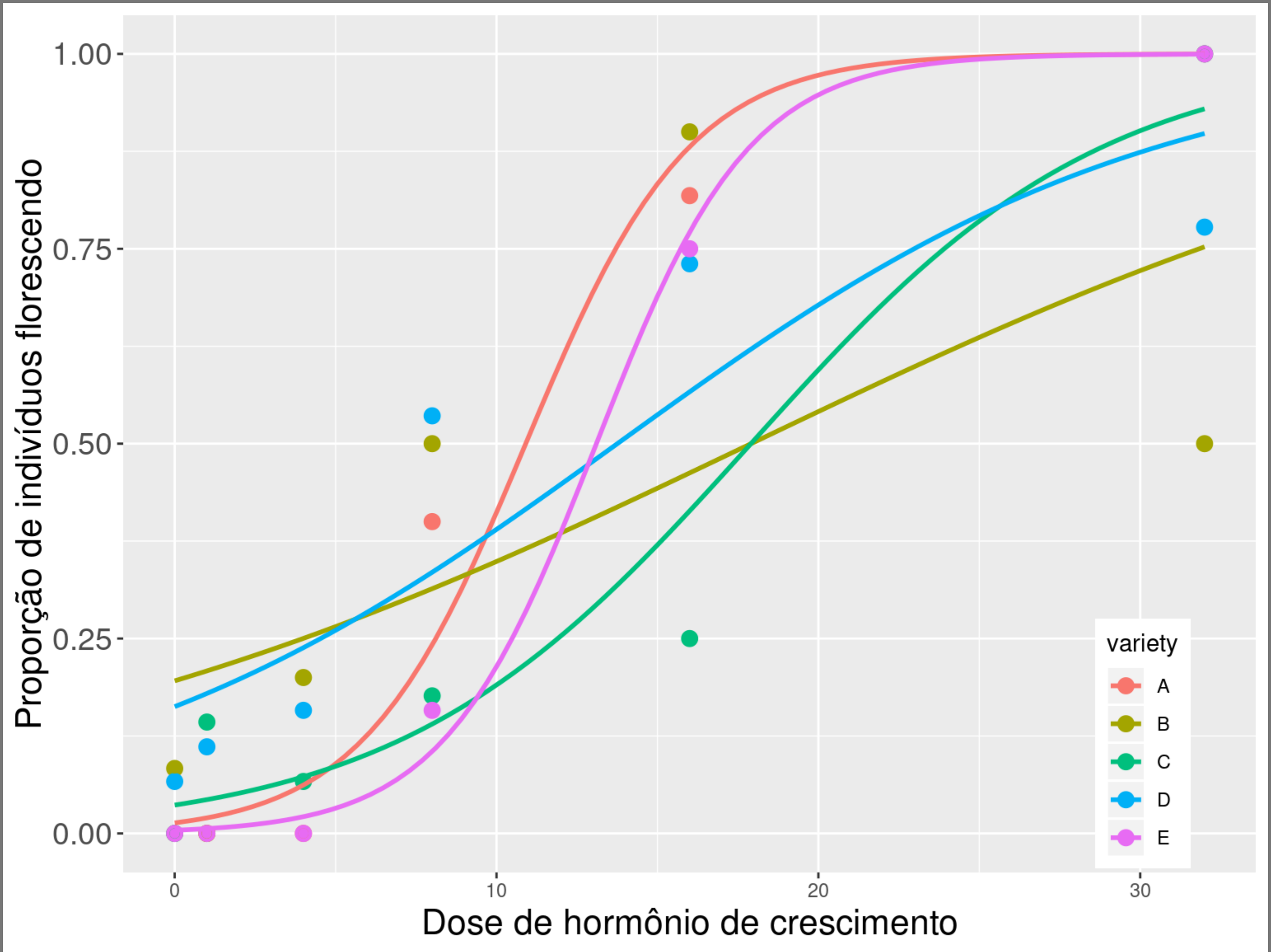
```
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

GLM Binomial: simplificando o modelo

```r
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


Proporção de indivíduos florescendo
1.00
0.75
0.50
0.25
0.00
0
10
20
30
Dose de hormônio de crescimento
variety
A
B
C
D
E

Predito pelo modelo: linear

```r
predict(bin01)
```

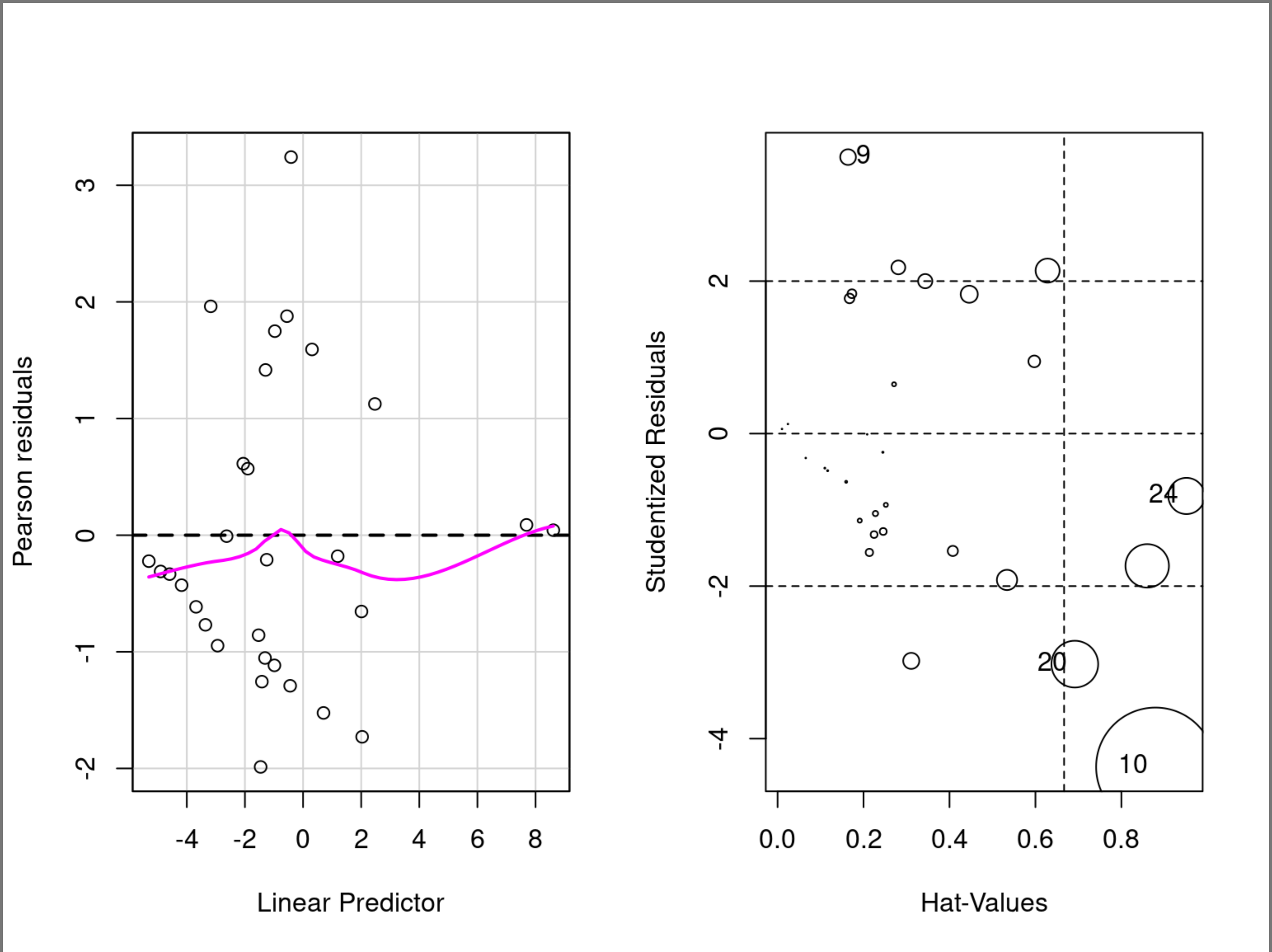
```
##          1          2          3          4          5
## -4.1790352 -2.9411862 -1.2907208  2.0102098  8.6120711 -1.45988
##          7          8          9         10         11
## -1.2504982 -0.9713115 -0.4129381  0.7038087 -3.1769379 -2.63024
##         13         14         15         16         17
## -1.9013292 -0.4434907  2.4721863 -1.3087930 -0.9853868 -0.55417
##         19         20         21         22         23
##  0.3082377  2.0330705 -4.9001833 -3.6818046 -2.0572996  1.19171
##         25         26         27         28         29
##  7.6897302 -4.5916515 -1.5296848 -3.3591677 -1.4165950 -5.30630
```

Predito pelo modelo: antilogit

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

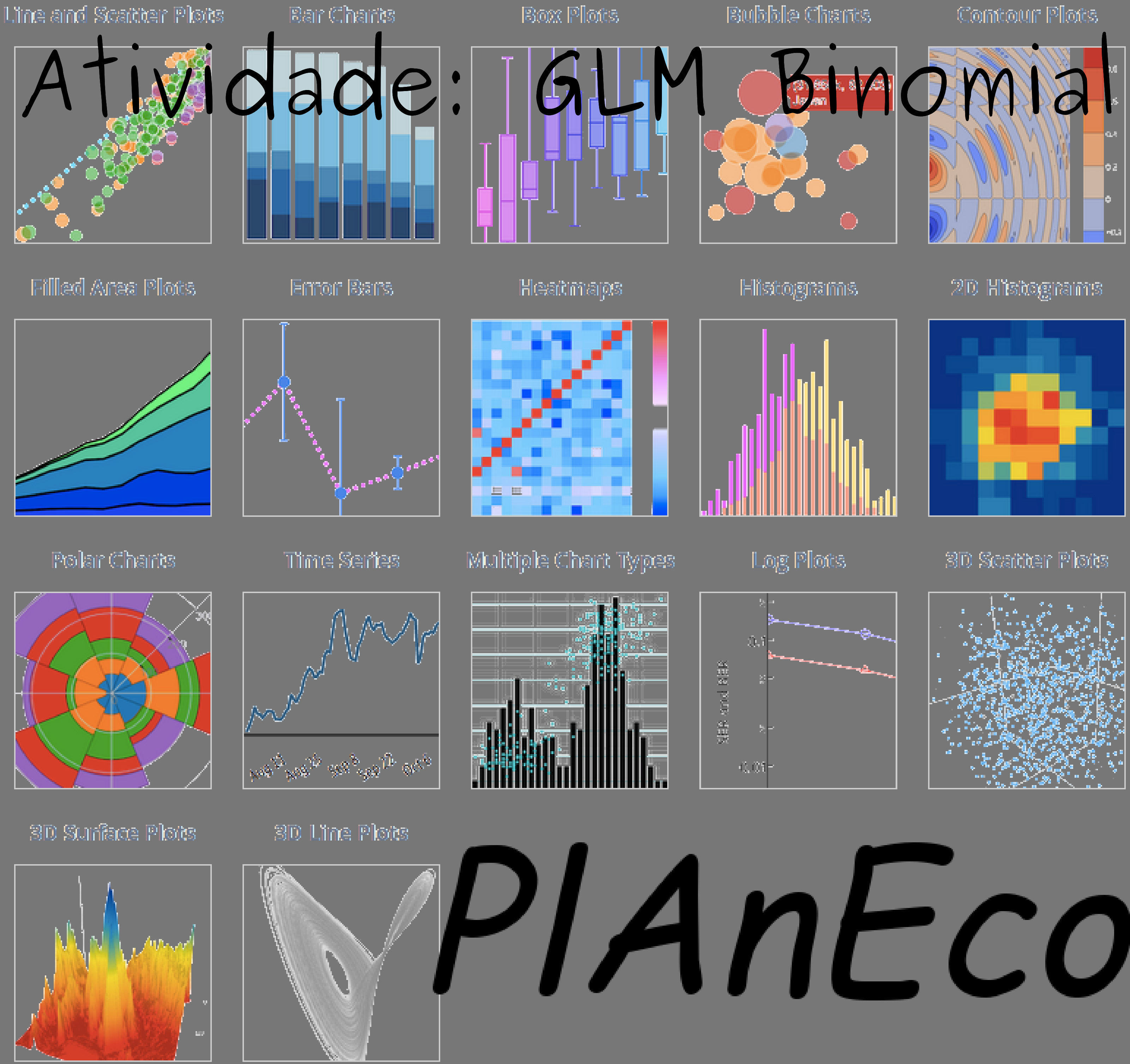
```
##           1           2           3           4           5
## 0.015082316 0.050154735 0.215730827 0.881864884 0.999818136 0.1
##           7           8           9          10          11
## 0.222613918 0.274619176 0.398207832 0.669031659 0.040042874 0.0
##          13          14          15          16          17
## 0.129958109 0.390909521 0.922168829 0.212688893 0.271824227 0.3
##          19          20          21          22          23
## 0.576455053 0.884225776 0.007390197 0.024559161 0.113316872 0.7
##          25          26          27          28          29
## 0.999542708 0.010034395 0.178039802 0.033596235 0.195195932 0.0
```

Diagnóstico do modelo



```
##        StudRes       Hat      CookD
## 9   3.6262074 0.1640033  0.2465669
## 10 -4.3712682 0.8793205 14.0234875
## 20 -3.0231527 0.6916136  2.1733865
## 24 -0.8177008 0.9513937  1.3097880
```

Atividade: GLM Binomial



PlAnEco

Exemplo: pássaros na ilha

incidence	area	isolation
1	7.93	3.32
0	1.92	7.55
1	2.04	5.88
0	4.78	5.93
0	1.54	5.31
1	7.37	4.93
1	8.60	2.88
0	2.42	8.77
1	6.40	6.09
1	7.20	6.98
0	2.65	7.75
1	4.13	4.30
0	4.17	8.52
1	7.10	3.32
0	2.39	9.29

Modelo Binomial: pássaros

```r
glmave <- glm(incidence~ area + isolation + area:isolation,
              family=binomial, data = ave)
summary(glmave)
```

```
##
## Call:
## glm(formula = incidence ~ area + isolation + area:isolation,
##     family = binomial, data = ave)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.84481  -0.33295   0.02027   0.34581   2.01591
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      4.0313     7.1747   0.562    0.574
## area             1.3807     2.1373   0.646    0.518
## isolation       -0.9422     1.1689  -0.806    0.420
## area:isolation  -0.1291     0.3389  -0.381    0.703
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 68.029  on 49  degrees of freedom
## Residual deviance: 28.252  on 46  degrees of freedom
## AIC: 36.252
```

```
## AIC: 30.232

##
## Number of Fisher Scoring iterations: 7
```

Modelo Binomial: pássaros

```r
glmave01<- glm(incidence~ area + isolation,
               family=binomial, data = ave)
anova(glmave01, glmave, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: incidence ~ area + isolation
## Model 2: incidence ~ area + isolation + area:isolation
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        47     28.402
## 2        46     28.252  1  0.15043   0.6981
```

Modelo Binomial: pássaros

```r
summary(glmave01)
```

```
##
## Call:
## glm(formula = incidence ~ area + isolation, family = binomial,
##     data = ave)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8189  -0.3089   0.0490   0.3635   2.1192
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   6.6417     2.9218   2.273  0.02302 *
## area          0.5807     0.2478   2.344  0.01909 *
## isolation    -1.3719     0.4769  -2.877  0.00401 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 68.029  on 49  degrees of freedom
## Residual deviance: 28.402  on 47  degrees of freedom
## AIC: 34.402
##
```

```
## Number of Fisher Scoring iterations: 6
```

Modelo Binomial: pássaros

```r
glmave02<- glm(incidence~ area,
               family=binomial, data = ave)


anova(glmave02, glmave01, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: incidence ~ area
## Model 2: incidence ~ area + isolation
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        48     50.172
## 2        47     28.402  1    21.77 3.073e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Modelo Binomial: pássaros

```
glmave03<- glm(incidence~ isolation,
               family=binomial, data = ave)
anova(glmave03, glmave01, test= "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: incidence ~ isolation
## Model 2: incidence ~ area + isolation
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        48     36.640
## 2        47     28.402  1   8.2375 0.004103 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

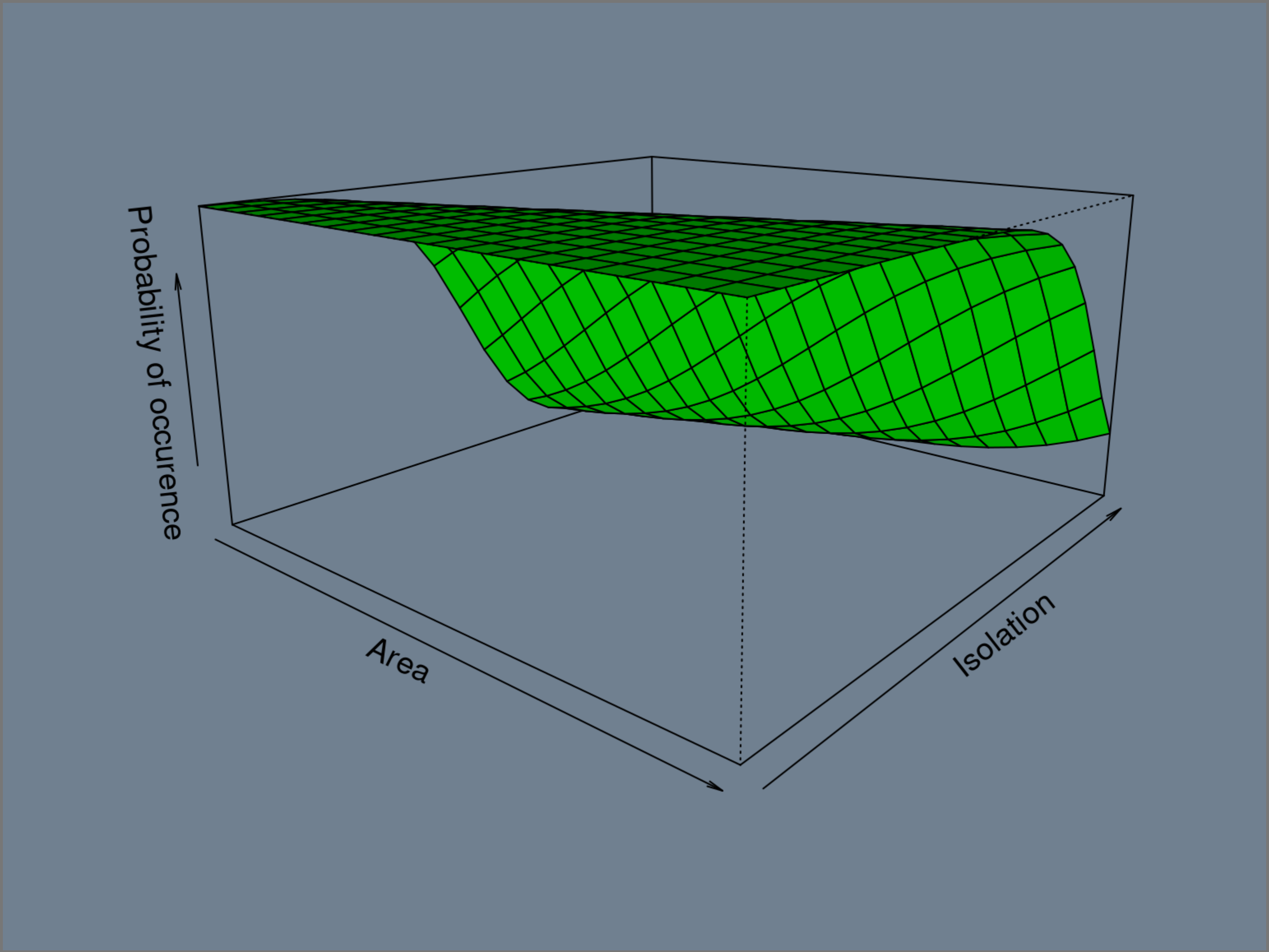
Modelo Binomial: pássaros

```r
summary(glmave01)
```

```
##
## Call:
## glm(formula = incidence ~ area + isolation, family = binomial,
##     data = ave)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8189  -0.3089   0.0490   0.3635   2.1192
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   6.6417     2.9218   2.273  0.02302 *
## area          0.5807     0.2478   2.344  0.01909 *
## isolation    -1.3719     0.4769  -2.877  0.00401 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 68.029  on 49  degrees of freedom
## Residual deviance: 28.402  on 47  degrees of freedom
## AIC: 34.402
##
```

```
## Number of Fisher Scoring iterations: 6
```

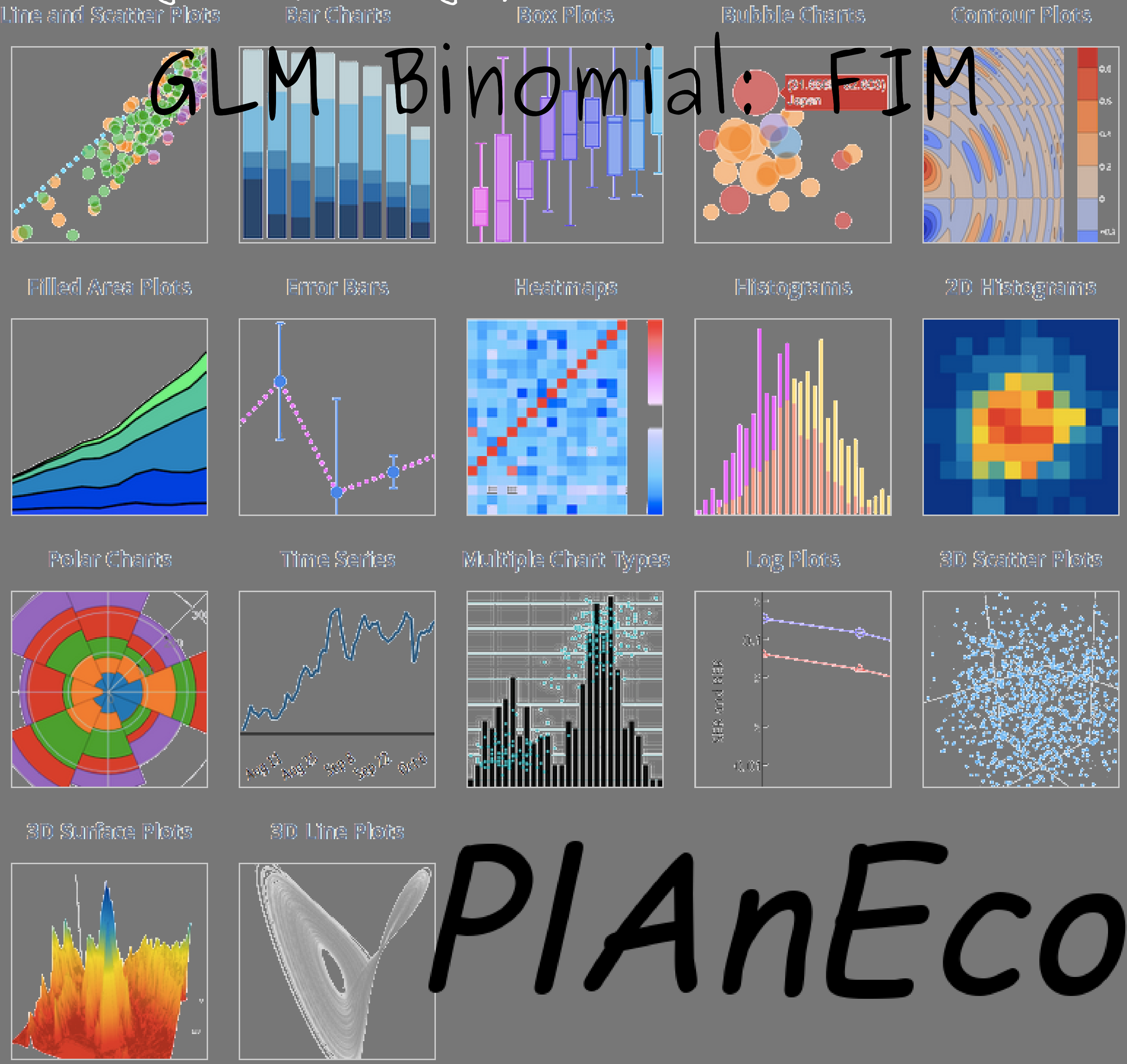
Gráfico do modelo



GLM: ajuste de modelo proporção

- ajuste a variável resposta (sucesso, falha) ou proporção
- não transforme (sucesso,falha) em proporção a menos que o número de tentativas seja igual para todas as unidades amostrais
- use a familia de ligação binomial(logit)
- avalie o sobre-dispersão do erro pela razão Residual deviance / degrees of freedom
- razão > 1, use família quasibinom
- busque o modelo mínimo adequado com anova
  - binomial use test = "Chisq"
  - quasibinom use test = "F"
- retenha o modelo mínimo adequado
- retorne os coeficientes e preditos do modelo para escala original (antilogit)

GLM Binomial: FIM



PlAnEco