Modelos lineales y aditivos en ecología

Facundo X. Palacio

facundo_palacio@fcnym.unlp.edu.ar

2 al 6 de mayo de 2022 – Universidad Nacional de Tucumán

Tipos de datos

- Conteos (GLM Poisson, binomial negativo, modelos log-lineales, inflados en ceros)
- Proporciones discretas, presencia-ausencia (GLM binomial)
- Datos continuos (GLM gamma)

¿Qué son los GLM?

- Nelder y Wedderburn (1972)
- Unificación de distintos modelos:

Regresión lineal

Regresión logística

Regresión de Poisson

Análisis de supervivencia…

Ventajas de los GLM

- Modelos paramétricos clásicos → Normalidad de residuos y homogeneidad de varianzas

- GLM:

- Distribuciones no normales!
- Varianzas no constantes!

¿Por qué no transformar?

- Los GLMs describen los datos en la escala original
- Transformaciones = interpretación biológica complicada
- Permiten trabajar con distribuciones no normales y varianzas no constantes

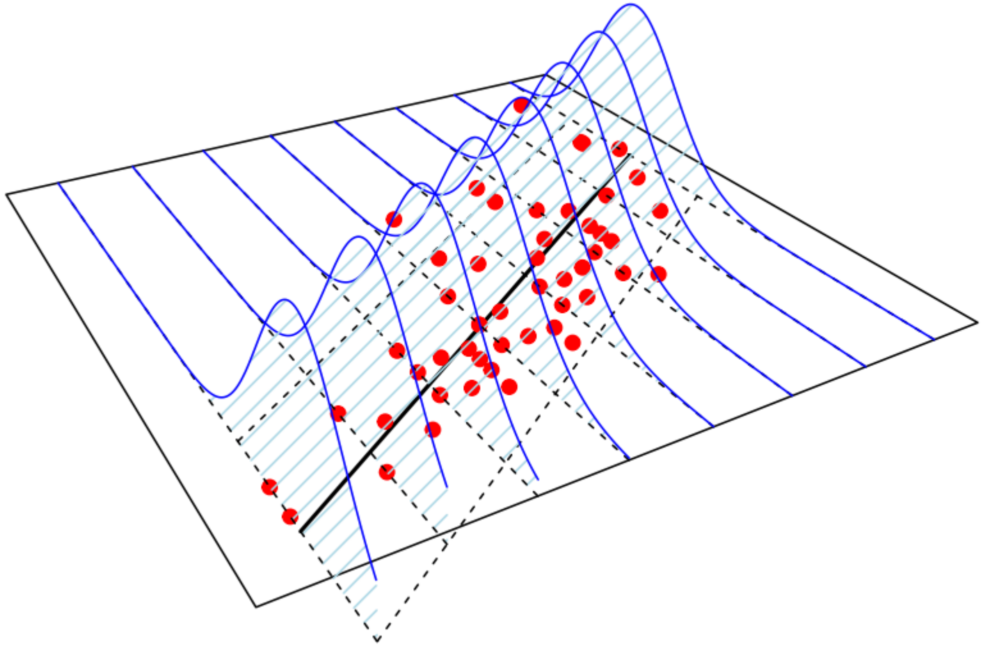
Un mismo marco teórico

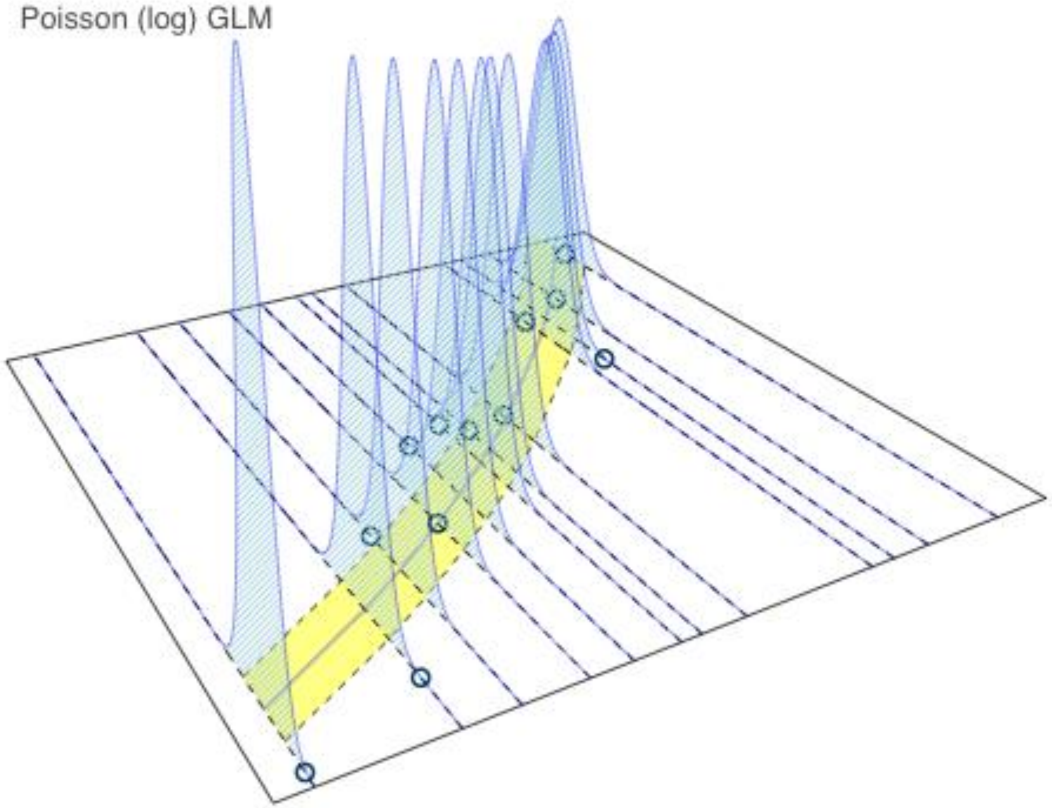
- Numerosas distribuciones de la variable dependiente (normal, Poisson, binomial) pertenecientes a un mismo *tipo* de distribución.
- Estimación de parámetros: *máxima verosimilitud*.
- Significancia de parámetros: *devianza* y test de t.

Componentes de un GLM

- 1. Estructura del error
- 2. Predictor lineal
- 3. Función de enlace

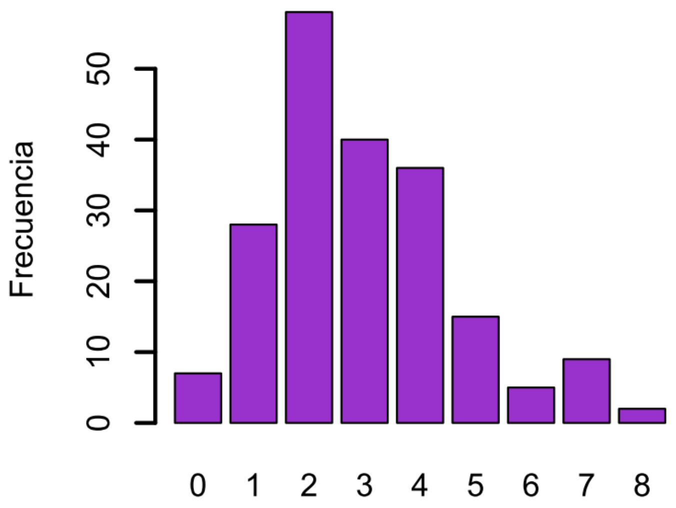
Estructura del error

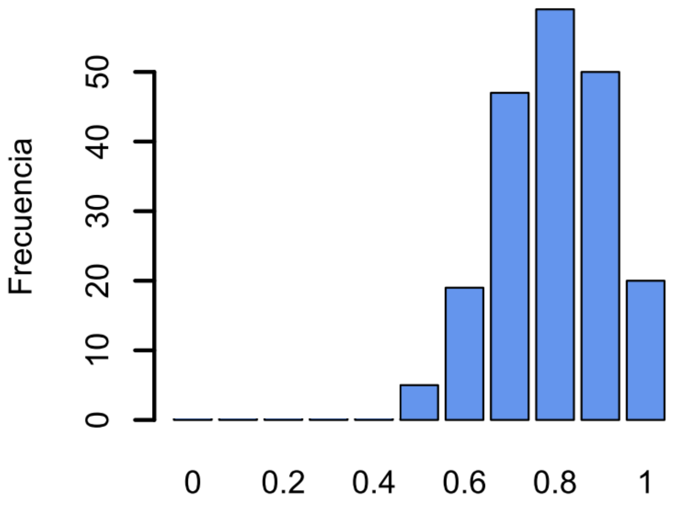


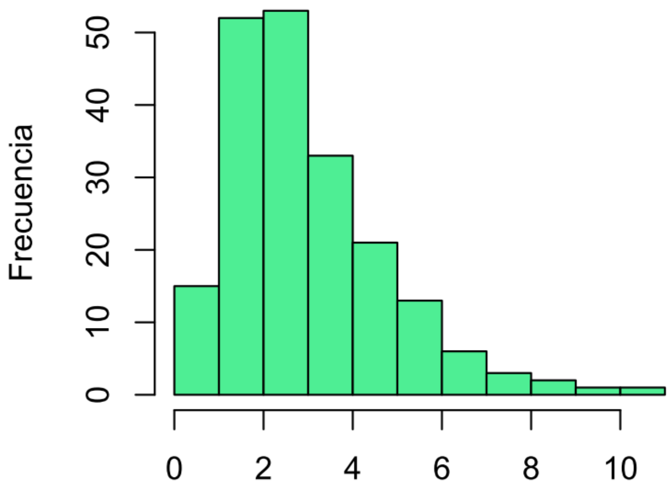


Estructura del error

- Distribución de la variable respuesta







Estructura del error

Naturaleza de los datos

N° de huevos/hembra

- Discreta
- [0, +∞)

Poisson, binomial negativa

Presencia-ausencia de una especie

- Datos binarios
- {0, 1}

binomial

Masa corporal

- Continua
- (0, +∞)

Gamma

Estructura del error

- Familia exponencial

- Normal
- Poisson
- Binomial
- Binomial negativa
- Gamma
- Weibull
- Etc., etc.

Predictor lineal

$$y = b_0 + b_1 x$$

$$y = \frac{1}{b_0 + b_1 x}$$

$$y = e^{b_0 + b_1 x}$$

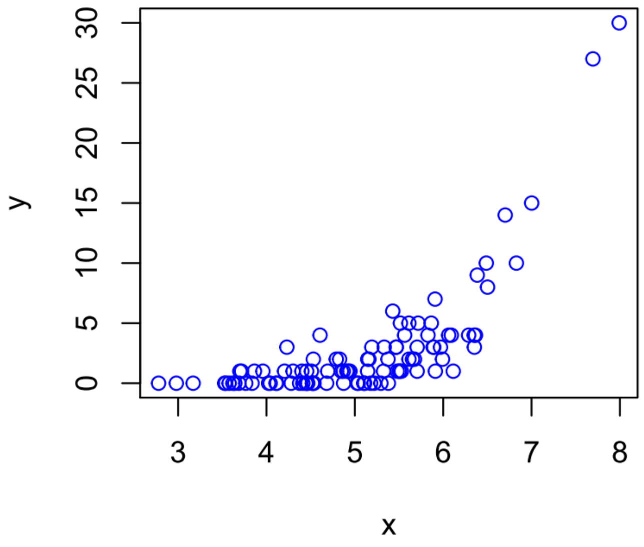
$$y = \frac{1}{\sqrt{b_0 + b_1 x}}$$

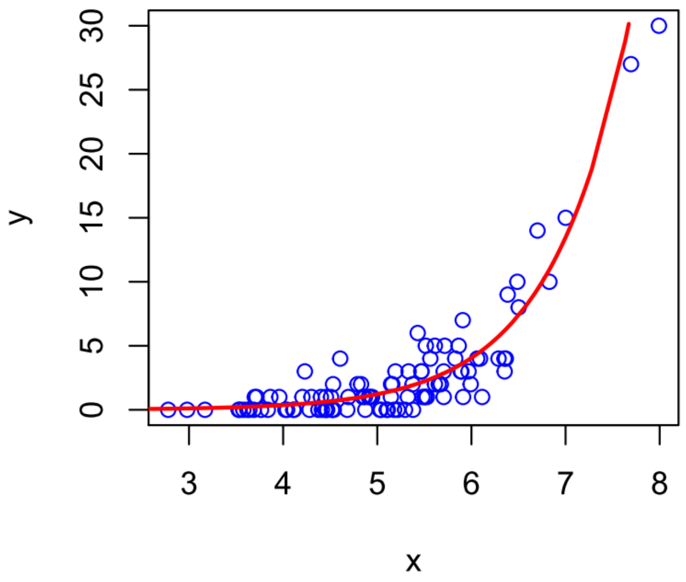
$$y = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$$

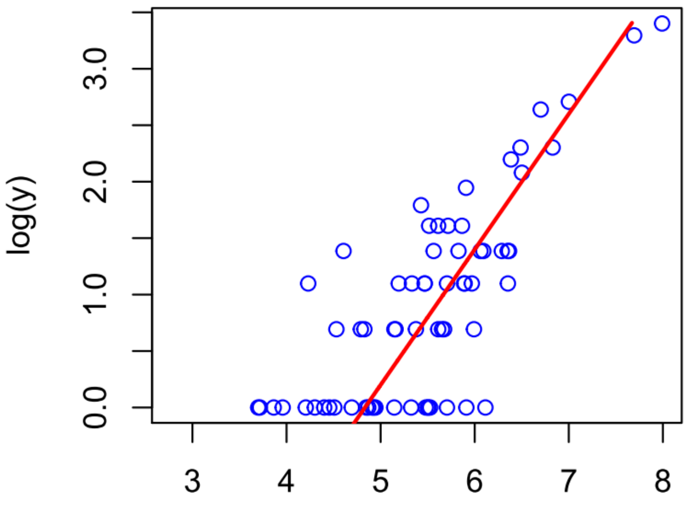
Función de enlace

- Relaciona la **estructura del error** con el **predictor lineal**

$$f\left(\hat{y}\right)=b_0+b_1x \qquad \hat{y}=e^{b_0+b_1x} \qquad \log\left(\hat{y}\right)=b_0+b_1x$$







Supuestos

- El modelo es lineal (una vez transformado).
- y tiene una distribución de probabilidad.
- Independencia de las observaciones.

- ¿Dónde están los residuos?
- ¿Dónde está la homogeneidad de varianzas?

Máxima verosimilitud

Fisher (1912-1922)

$$x = 2,0,5,3,1,3,0,4,5,4$$

$$\bar{x} = 2.7$$

$$H_0 : \mu = 0$$

$$H_1 : \mu \neq 0$$

Intervalos de confianza

Máxima verosimilitud

$$x = 2,0,5,3,1,3,0,4,5,4$$

- ¿Cuál es la media más probable de la población?

1. Calcular la probabilidad de la muestra tomando distintos valores de media poblacional.
2. Quedarse con el valor de media que maximiza la probabilidad de obtener los datos observados

Máxima verosimilitud

$$x = 2,0,5,3,1,3,0,4,5,4$$

$$P(muestra) = P(2)\times P(0)\times P(5)\times P(3)\times P(1)\times P(3)\times P(0)\times P(4)\times P(5)\times P(4)$$

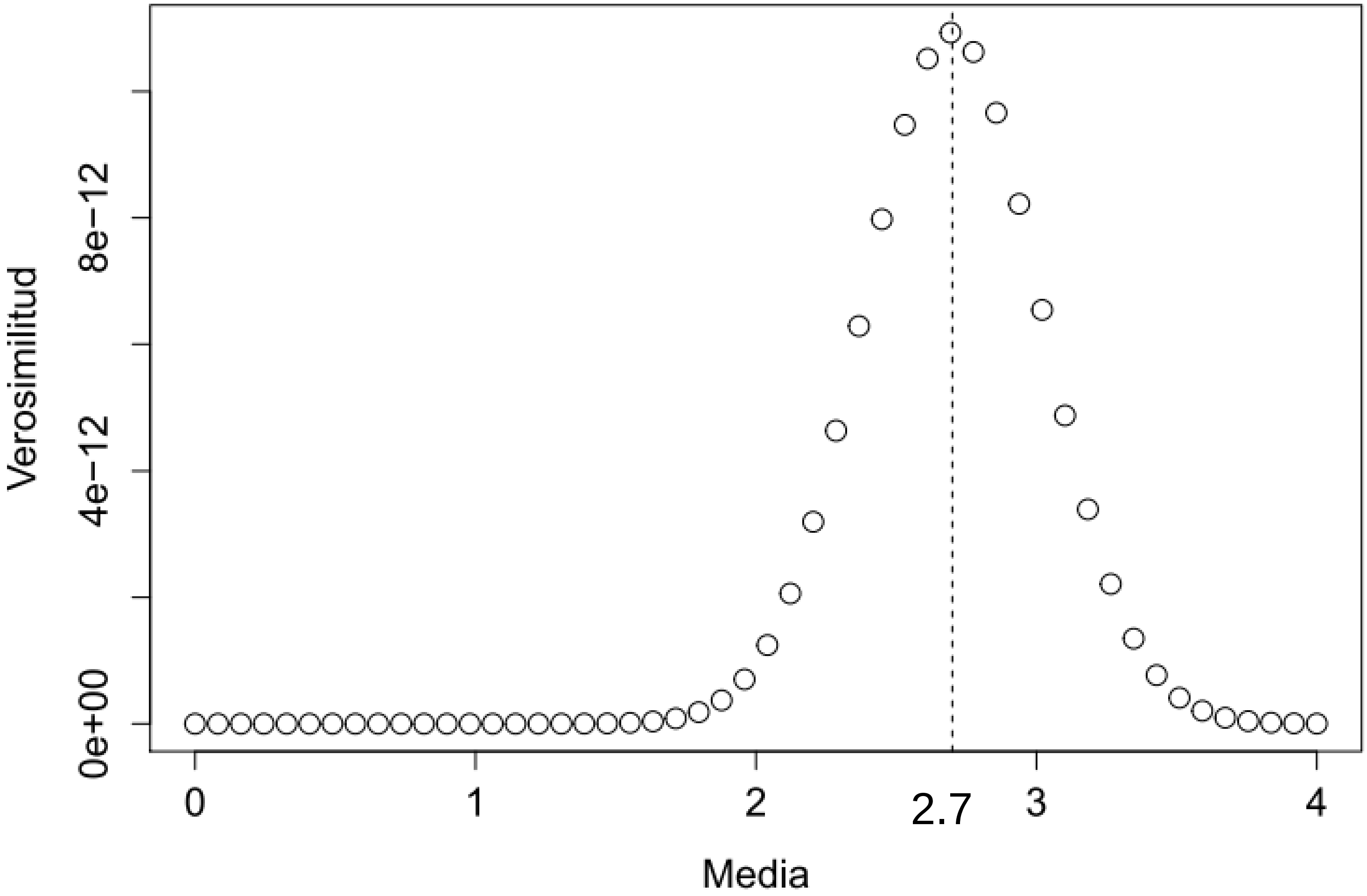
$$= \prod P(x_i)$$

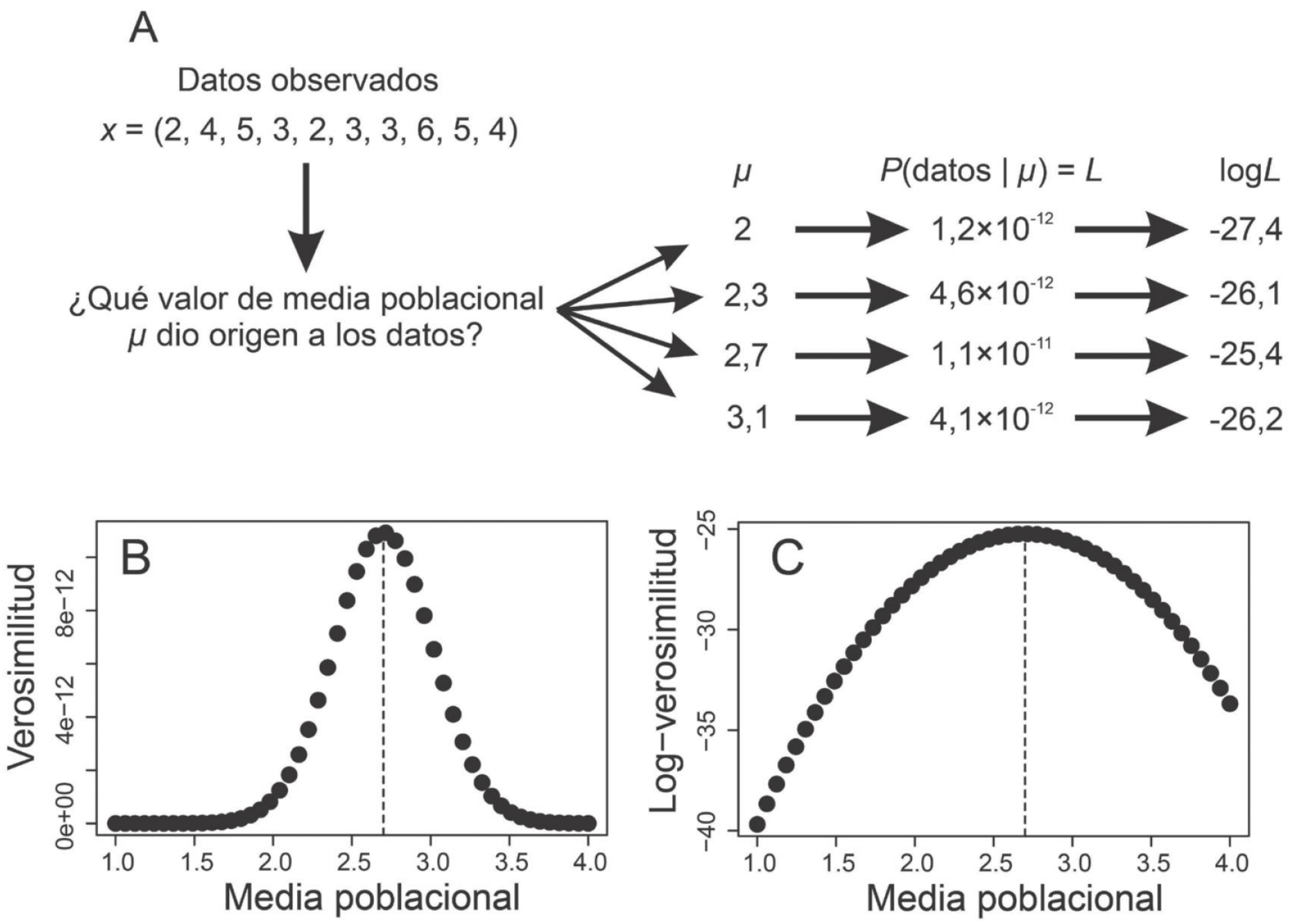
Distribución de probabilidad

Probar distintos valores del parámetro (e.g. media)

$$P(muestra \mid parámetro) = L$$

Máxima verosimilitud




A
Datos observados
x = (2, 4, 5, 3, 2, 3, 3, 6, 5, 4)
¿Qué valor de media poblacional μ dio origen a los datos?
μ
P(datos | μ) = L
logL
2
1,2×10^-12
-27,4
2,3
4,6×10^-12
-26,1
2,7
1,1×10^-11
-25,4
3,1
4,1×10^-12
-26,2
B
Verosimilitud
0e+00
4e−12
8e−12
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Media poblacional
C
Log−verosimilitud
−40
−35
−30
−25
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Media poblacional

Máxima verosimilitud

$$\hat{y} = b_0 + b_1 x$$

- ¿Cuáles son los valores de b_0 y b_1 que **maximizan** la probabilidad de obtener los datos observados?

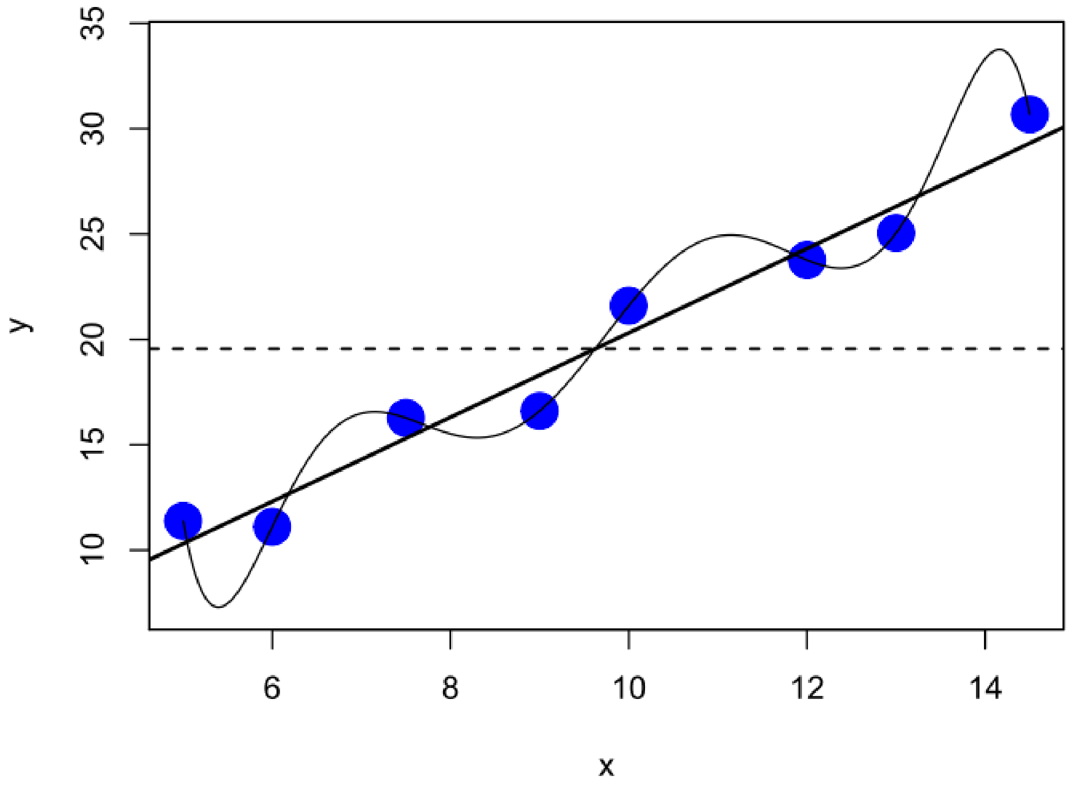
Devianza

- **Modelo candidato**

$$\hat{y} = b_0 + b_1 x$$

- **Modelo saturado**

$$\hat{y} = b_0 + b_1 x + \ldots + b_7 x^7$$



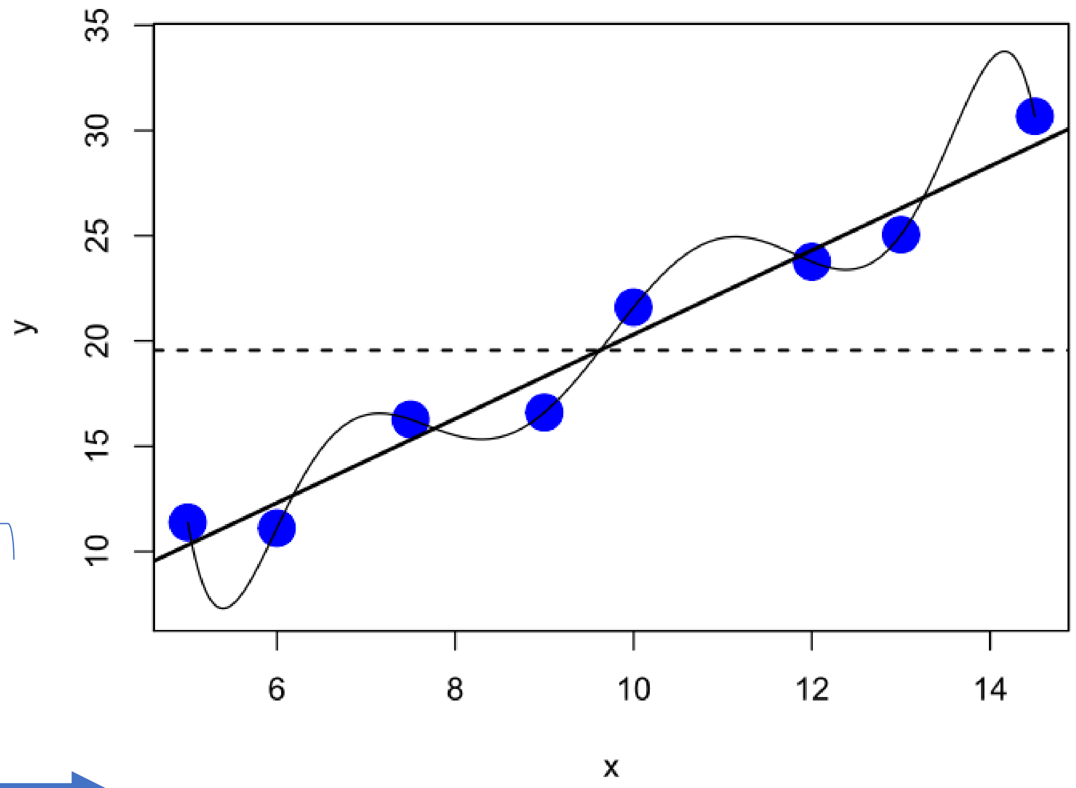
$$\hat{y} = b_0$$

$$D = -2\log\left(\frac{L \text{ modelo candidato}}{L \text{ modelo saturado}}\right)$$

$$= -2\left(\log L \text{ modelo candidato} - \log L \text{ modelo saturado}\right)$$

Modelo nulo $D_0 = -2\left(\log L \text{ modelo nulo} - \log L \text{ modelo saturado}\right)$

Devianza



$$\overbrace{\hspace{6em}}^{D}$$

Modelo candidato

Modelo nulo (peor modelo)

Modelo saturado (ajuste perfecto)

$$\underbrace{\hspace{12em}}_{D_0}$$

$$\text{pseudo-}R^2 = \frac{D_0 - D}{D_0} = 1 - \frac{D}{D_0}$$

Tests de hipótesis

$$H_0 : \beta_1 = \beta_2 = ... = \beta_n = 0$$

$$H_1 : \text{al menos un } \beta_n \neq 0$$

$$t = \frac{b_i}{S_{b_i}}$$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.5794     3.1214   1.787    0.112
x             4.3700     0.5573   7.841 5.04e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Validación

- Tipos de residuos

- **Residuos crudos**

e = obs – esp

- **Residuos estandarizados, normalizados o de Pearson**

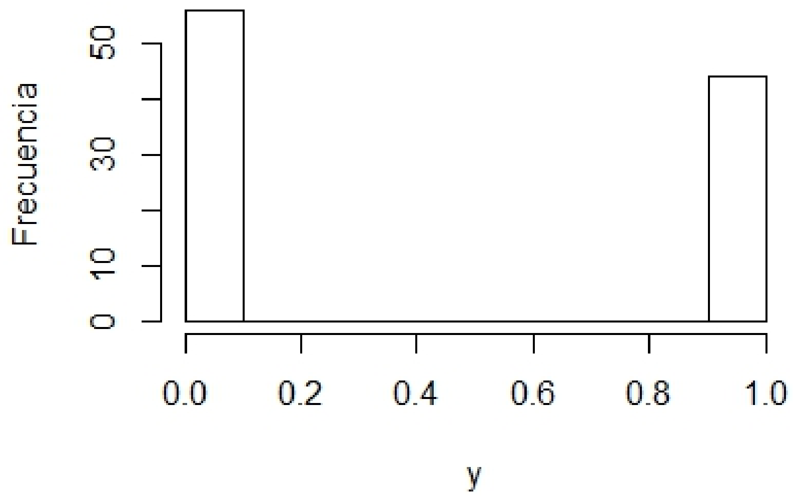
e = (obs – esp)/desvío

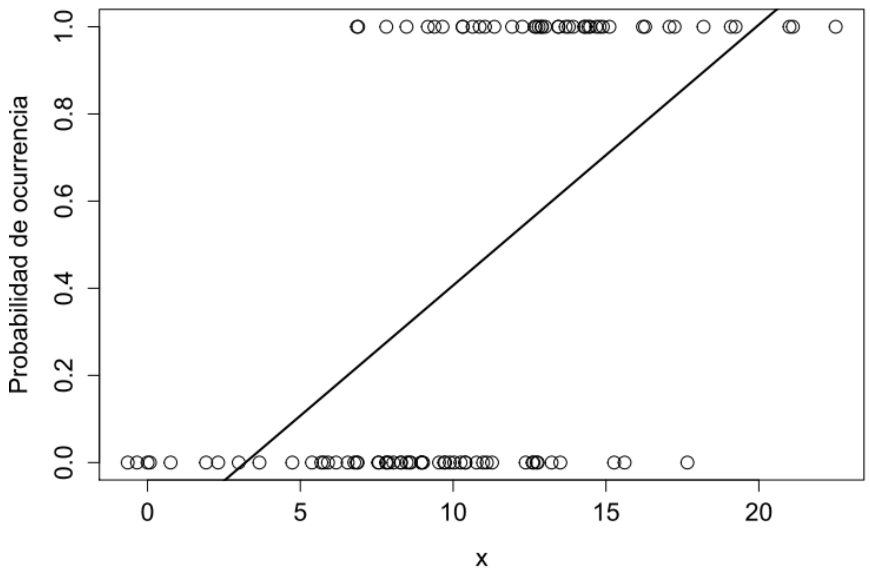
- **Residuos de devianza**

Cuánto contribuye un residuo a la devianza

Datos binarios: GLM binomial

• Datos binarios
• Propiedades:
- 2 valores: 0 y 1
- Distribución no normal
- Varianza no cte

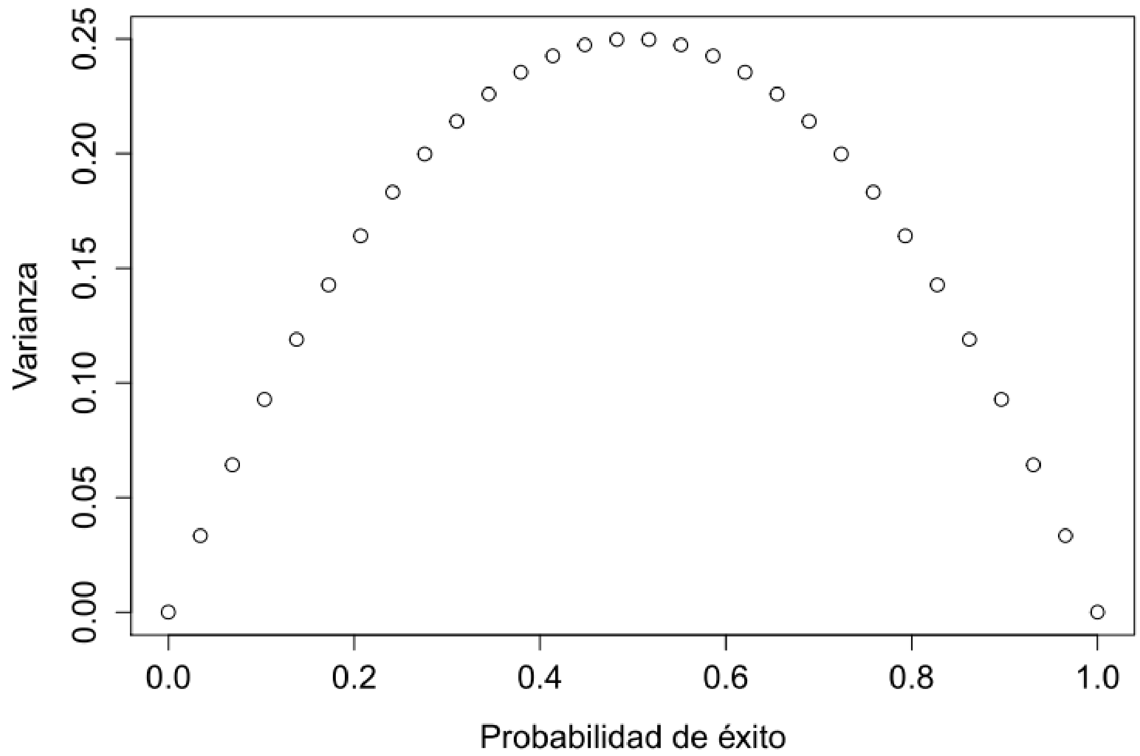




Componentes de un GLM binomial

- Distribución del error: Bernoulli
- Parámetro: p (probabilidad de éxito)

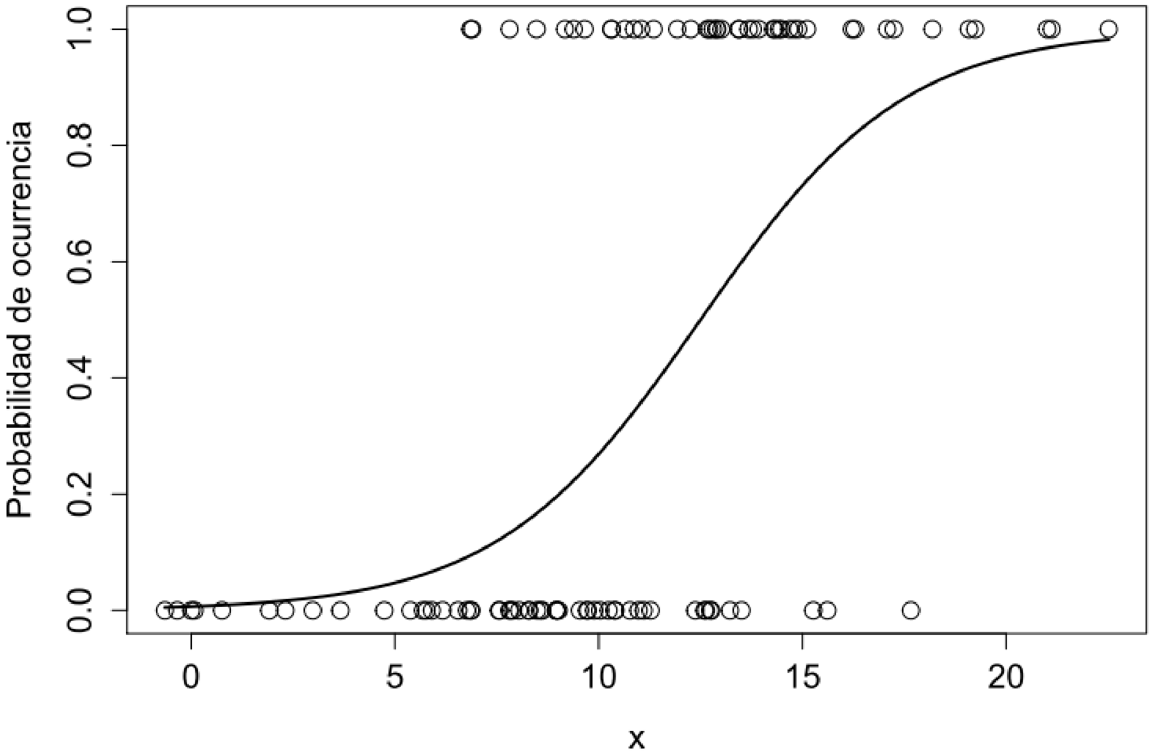
$\mathrm{Var}(p) = p(1 - p) = p - p^2$



Componentes de un GLM binomial

$$p = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$$

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1 x$$



Odds y logits

$$p = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$$

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1 x$$

$$\frac{p}{1-p} = e^{b_0 + b_1 x}$$

$$\text{razón de odds } = \frac{\text{odd}(x+1)}{\text{odd}(x)} = e^{b_1}$$

Odd

O = 8/2

10 veces, gana 8, pierde 2

Cambio en el odd de la respuesta por unidad de x

$b_1 > 0$ → el odd y p aumentan con x
$b_1 < 0$ → el odd y p disminuyen con x
$b_1 = 0$ → el odd y p es la misma para cada nivel de x

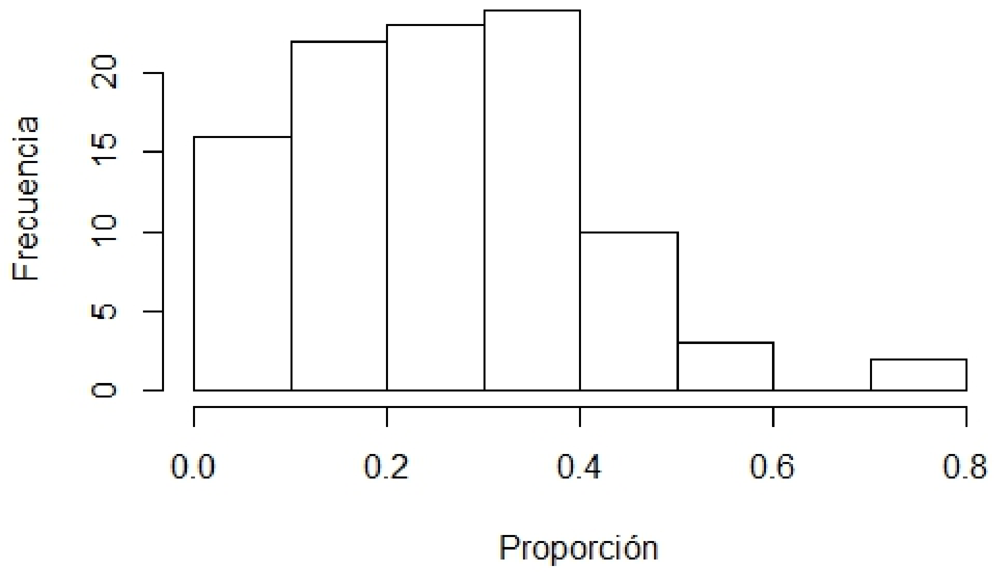
Proporciones: GLM binomial

Componentes

- Distribución del error: binomial
- Parámetros:

p (probabilidad de éxito)

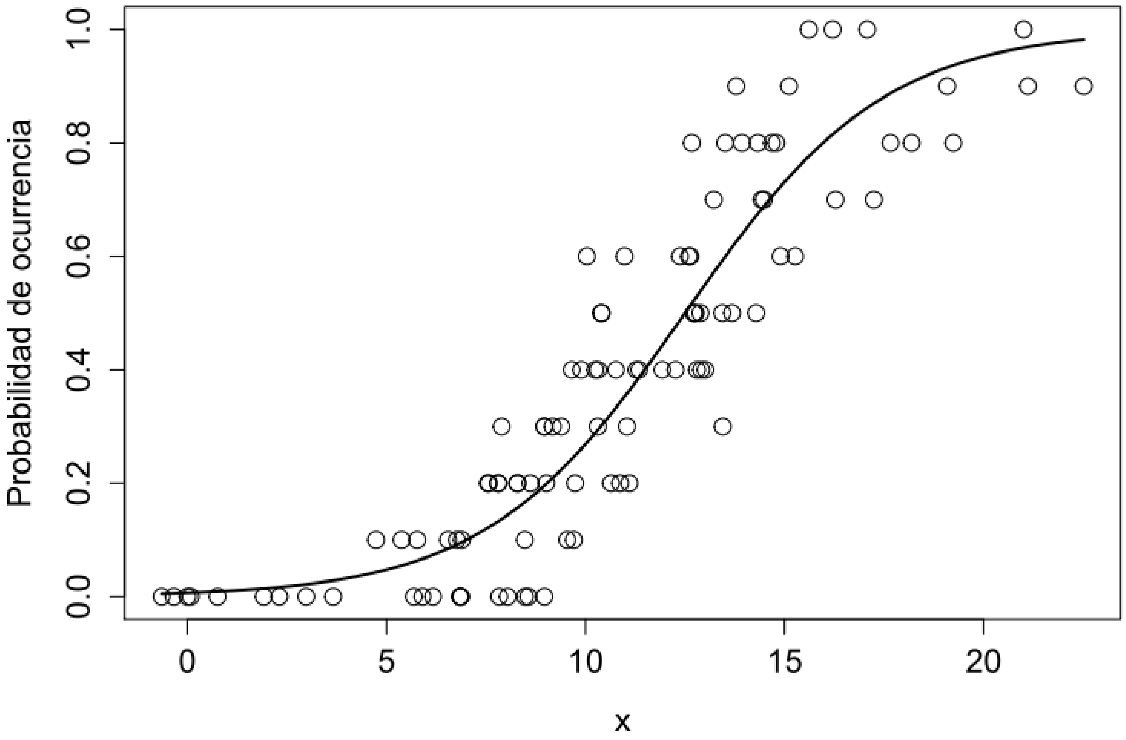
n (tamaño de muestra)



Proporciones: GLM binomial

$$p = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$$

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1 x$$



Bondad del ajuste

- R^2 de Tjur (2009)

Compara el promedio de las probabilidades predichas de los dos resultados posibles (0 y 1):

$$R^2_{Tjur} = \frac{\sum \hat{p}_1}{n_1} - \frac{\sum \hat{p}_0}{n_0}$$

Obs = 1, 1, 1, 1, 0, 0, 0

p = 0.7, 0.8, 0.6, 0.6, 0.1, 0.1, 0.2 → $R^2_{Tjur} = 0.67 - 0.13 = 0.54$

p = 1, 1, 1, 1, 0, 0, 0 → $R^2_{Tjur} = \frac{4}{4} - \frac{0}{3} = 1$

Variables discretas: GLM Poisson

- Conteos

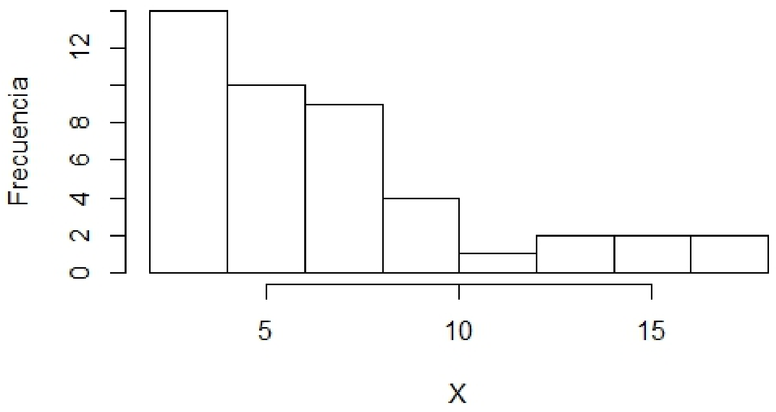
Propiedades:

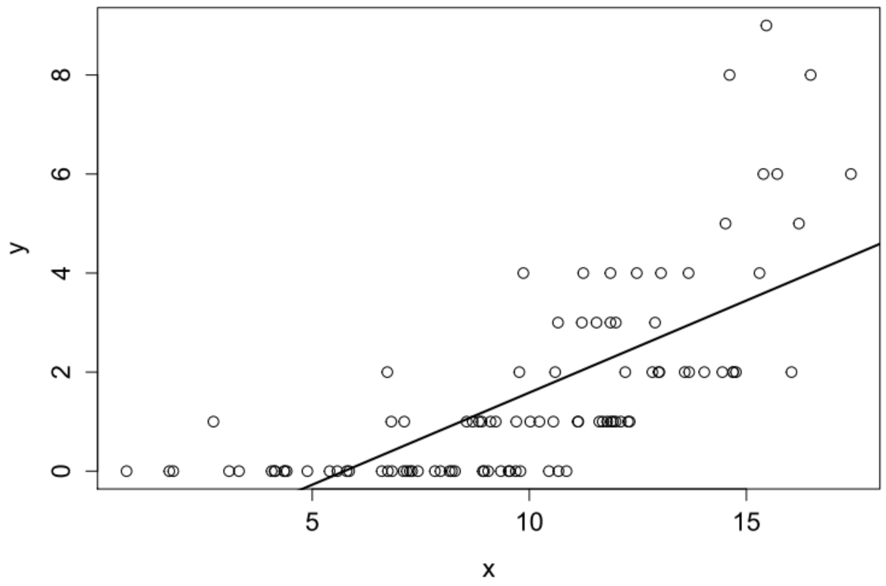
- Naturales (mínimo = 0)
- Distribución no normal

Componentes de un GLM Poisson

- Distribución del error
- Parámetro:

Media = Varianza

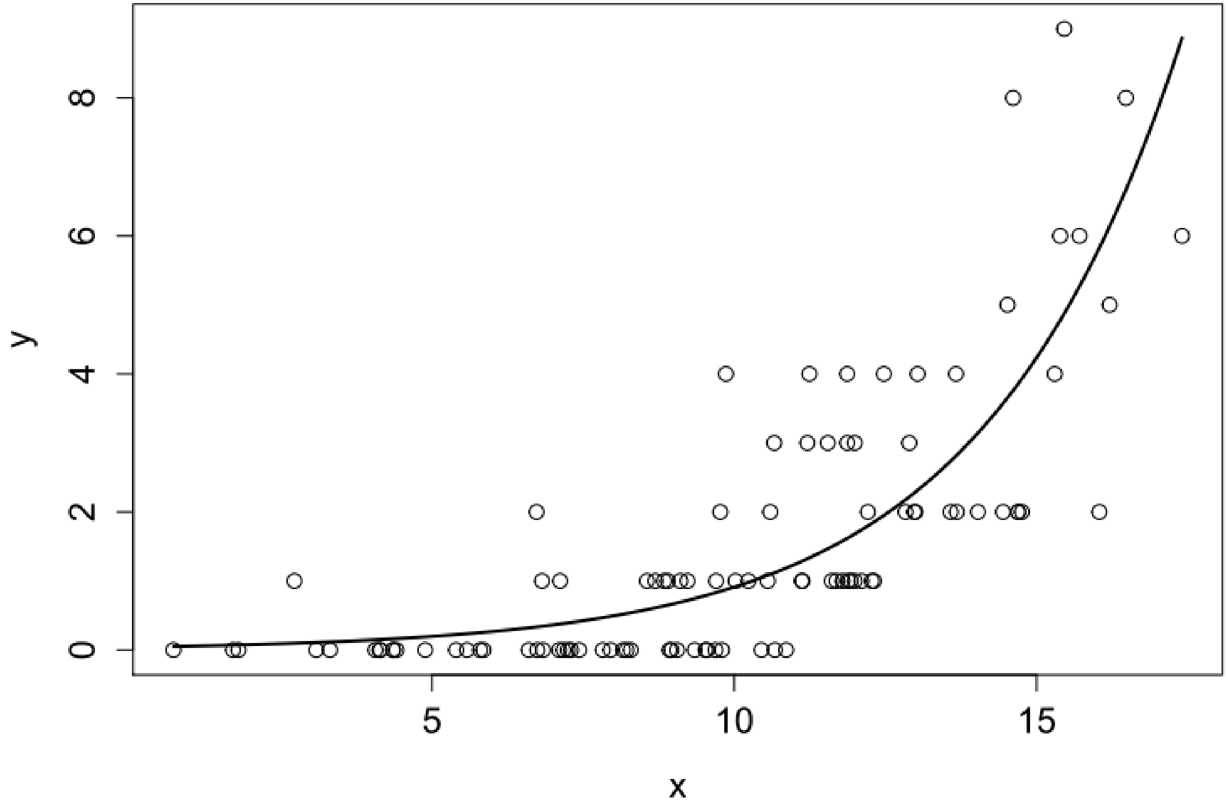




Componentes de un GLM Poisson

$$\hat{y} = e^{b_0 + b_1 x}$$

$$\log(\hat{y}) = b_0 + b_1 x$$

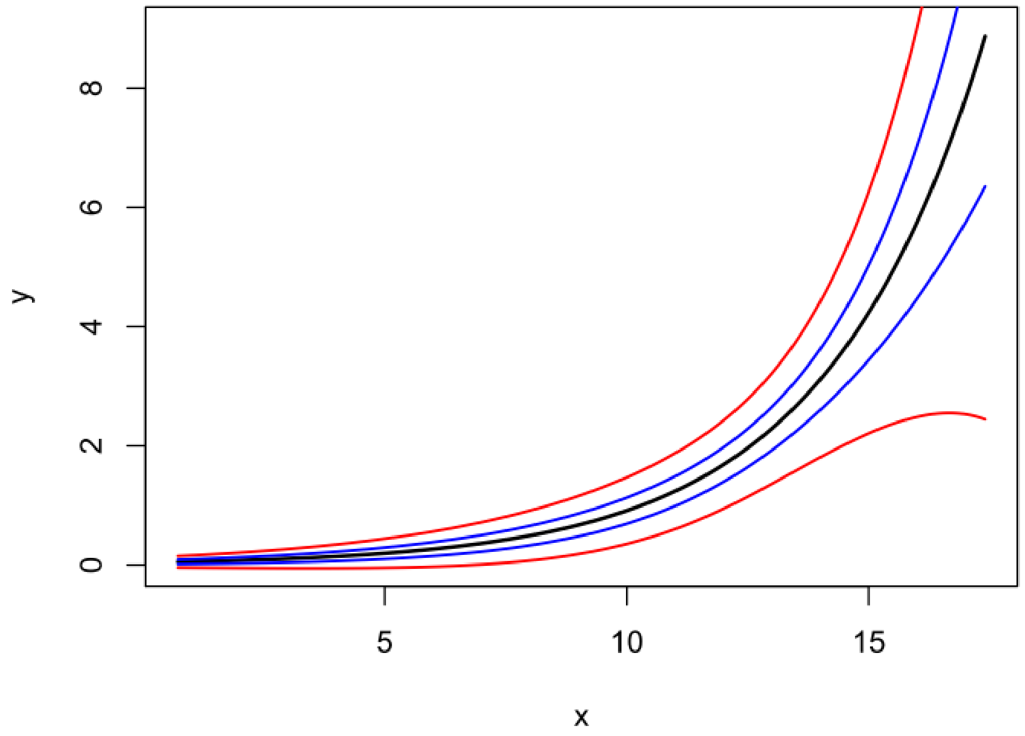


Sobredispersión

- Varianza > Media ➡ Inflan los valores de P

Posibles causas:

- No se incluyeron variables importantes
- Presencia de outliers
- Otra distribución de y
- Falta de independencia
- Excesos de ceros



Sobredispersión

- Solución 1)

Parámetro de sobredispersión

Media

Varianza = $\phi \times$ Media

GLM quasi-Poisson

Poisson vs quasi-Poisson

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)  3.054513   0.338925   9.012  < 2e-16 ***
typen       -0.814349   0.050379 -16.164  < 2e-16 ***
typeu       -0.277863   0.042799  -6.492 8.45e-11 ***
temp         0.018143   0.003768   4.815 1.47e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.05451    1.43017   2.136 0.040449 *  
typen       -0.81435    0.21259  -3.831 0.000562 ***
typeu       -0.27786    0.18060  -1.539 0.133742    
temp         0.01814    0.01590   1.141 0.262284    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 17.8062)
```

Validación

- Residuos normalizados (residuos de Pearson)

$$e = \frac{y - \hat{y}}{S} \Rightarrow \text{Sin sobredispersión}$$

$$e = \frac{y - \hat{y}}{\sqrt{\phi} S} \Rightarrow \text{Con sobredispersión}$$

Conteos II: GLM binomial negativo

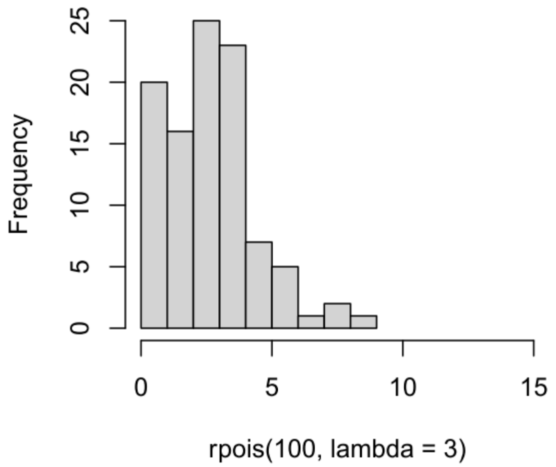
- Modelo alternativo al GLM Poisson y quasi-Poisson
- Componentes

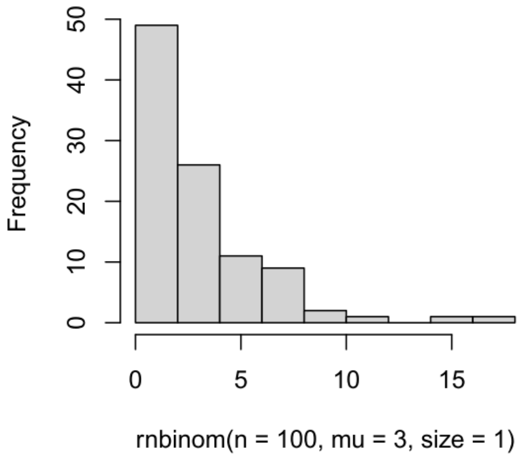
Distribución del error

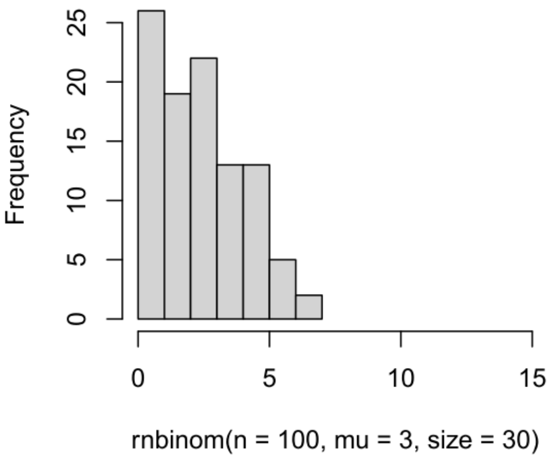
Varianza = Media + Media²/θ

$$\hat{y} = e^{b_0 + b_1 x}$$

$$\log\left(\hat{y}\right) = b_0 + b_1 x$$







Offset

- Conteos por *unidad* de superficie, volumen o tiempo

$$\frac{y}{n} = e^{b_0 + b_1 x}$$

$$y = n e^{b_0 + b_1 x}$$

$$\log(y) = \log(n) + b_0 + b_1 x = b'_0 + b_1 x$$

Offset

Offset

```
Call:
glm.nb(formula = total ~ habitat + temp + wind, data = pol, init.theta = 5.957692263,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3561  -0.8982  -0.1773   0.5760   1.8680

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.82455    1.28889   2.191  0.02842 *
habitata     0.78487    0.19397   4.046  5.2e-05 ***
habitatu     0.52773    0.18051   2.924  0.00346 **
temp         0.01218    0.01406   0.866  0.38652
wind        -0.01878    0.10041  -0.187  0.85161
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Call:
glm.nb(formula = total ~ habitat + temp + wind + offset(log(min)),
    data = pol, init.theta = 5.957692263, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3561  -0.8982  -0.1773   0.5760   1.8680

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.67526    1.28889  -1.300  0.19368
habitata     0.78487    0.19397   4.046  5.2e-05 ***
habitatu     0.52773    0.18051   2.924  0.00346 **
temp         0.01218    0.01406   0.866  0.38652
wind        -0.01878    0.10041  -0.187  0.85161
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Variables continuas: modelo lineal general

- Modelos paramétricos clásicos:

Test de t, ANOVAs varios, regresión…

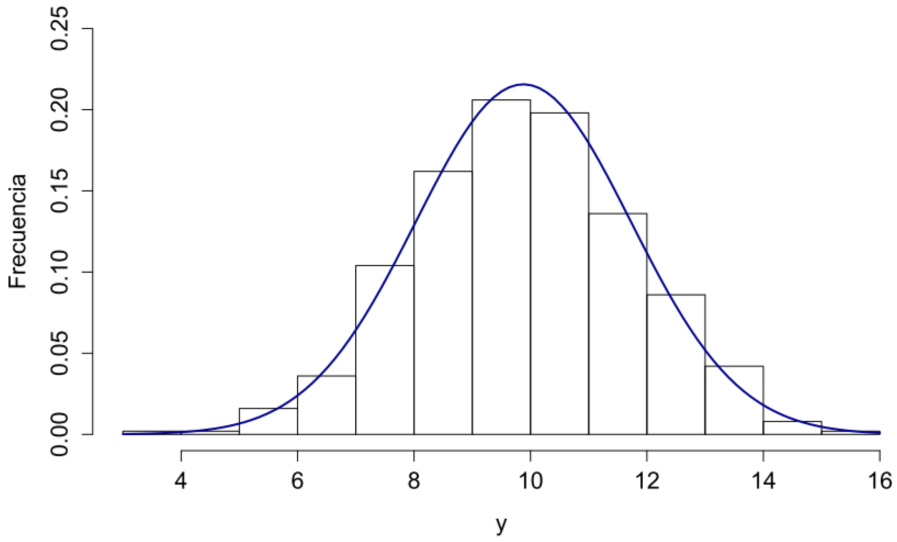
Componentes:

- Distribución del error

$$\hat{y} = b_0 + b_1 x$$

- ¿Función de enlace?

$$f(x) = x$$



Variables continuas: GLM gamma

- Datos continuos
- Propiedades:

- Valores positivos
- Varianza no cte

Variables continuas: GLM gamma

- Componentes

- Distribución del error

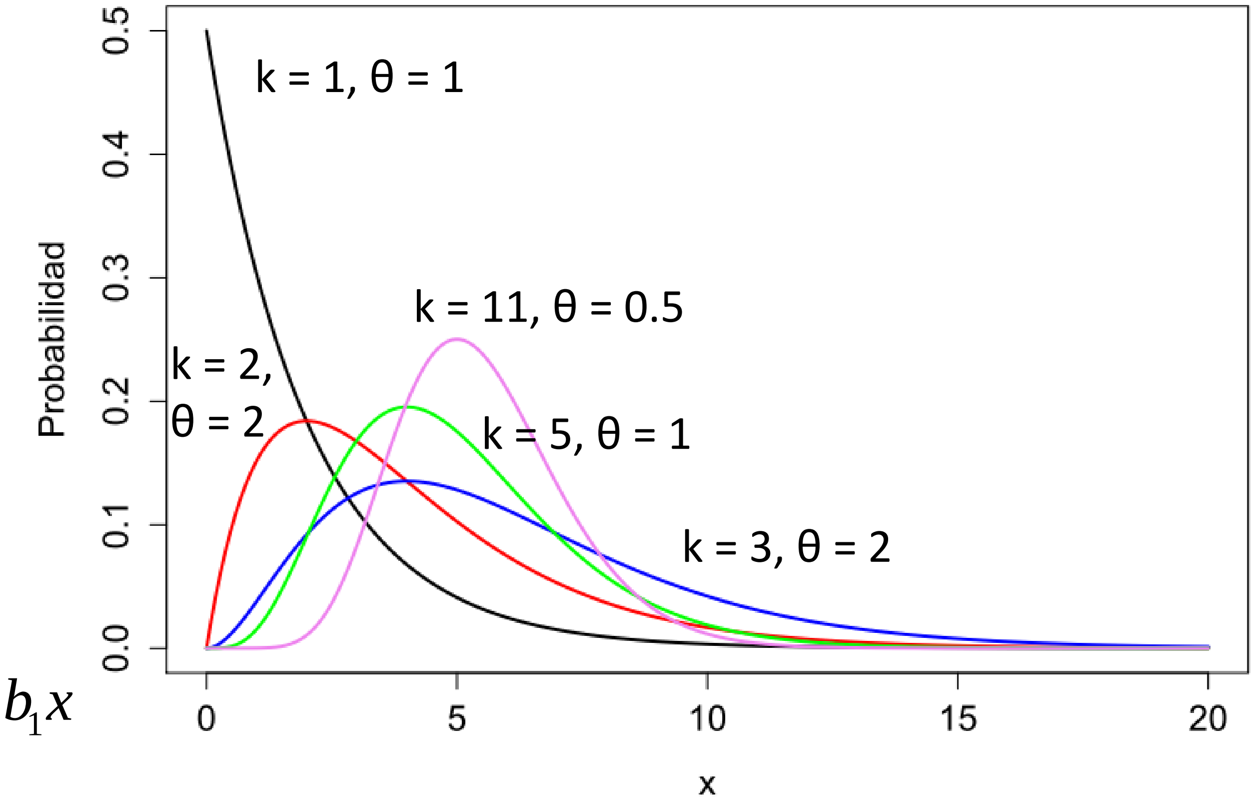
Parámetros:

k y θ

Var(Y) = μ²/k

$$\hat{y} = \frac{1}{b_0 + b_1 x}$$

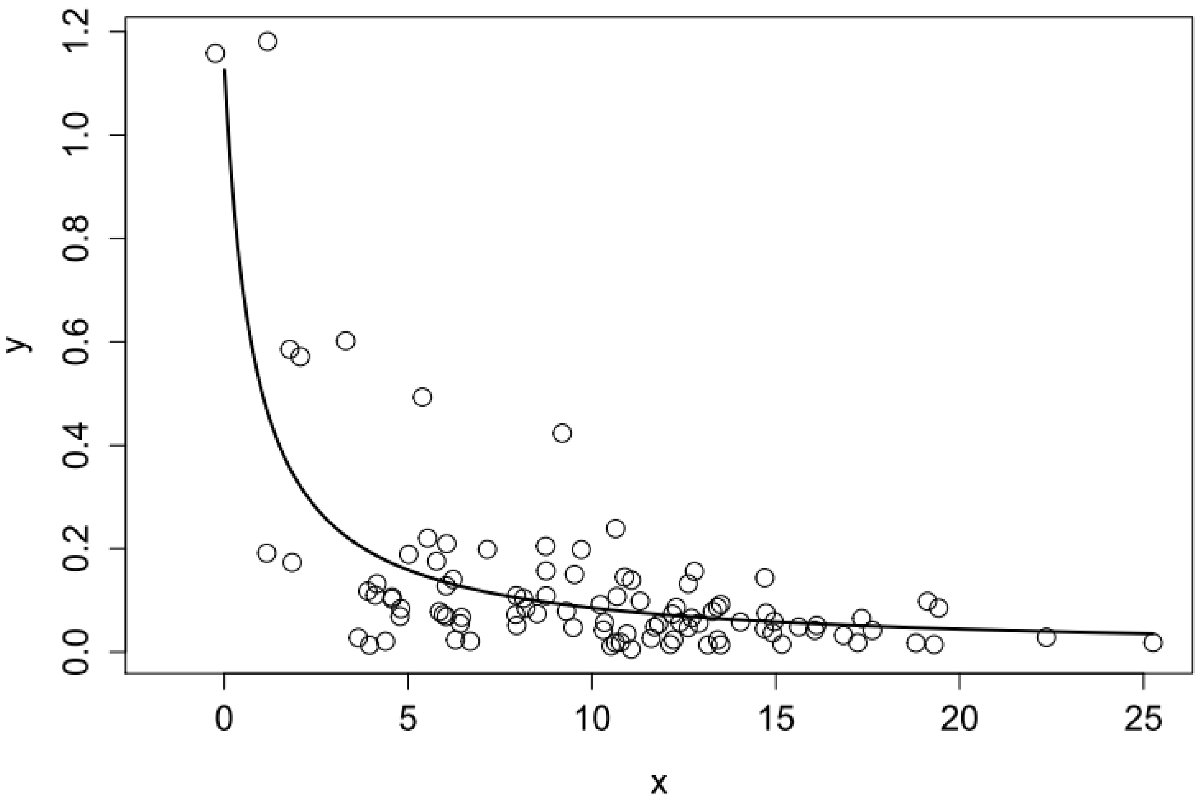
$$\frac{1}{\hat{y}} = \left(\frac{1}{b_0 + b_1 x}\right)^{-1} = b_0 + b_1 x$$



Variables continuas: GLM gamma

$$\hat{y} = \frac{1}{b_0 + b_1 x}$$

$$\frac{1}{\hat{y}} = \left(\frac{1}{b_0 + b_1 x}\right)^{-1} = b_0 + b_1 x$$




```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8775     0.4177   2.101   0.0383 *
x             1.0844     0.1068  10.149   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```