Modelos lineales y aditivos en ecología

Facundo X. Palacio

facundo_palacio@fcnym.unlp.edu.ar

CONICET

Facultad de Ciencias Naturales y Museo

SECCIÓN ORNITOLOGÍA · MUSEO DE LA PLATA · FCNYM - UNLP

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

Modelos truncados en cero

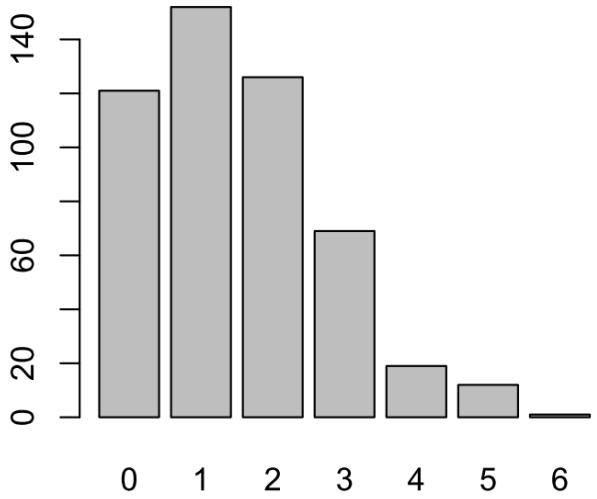
- Conteos

Propiedades:

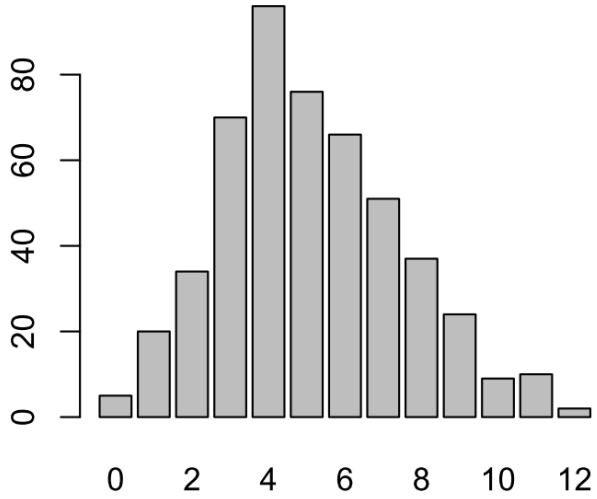
- Enteros positivos (mínimo = 1).
- Distribución no normal
- Distribución del error: Poisson, binomial negativa

Los modelos truncados en cero no son un problema, sino los supuestos de la distribución de conteos que admiten al 0 en su dominio → Excluimos al 0 de la distribución.

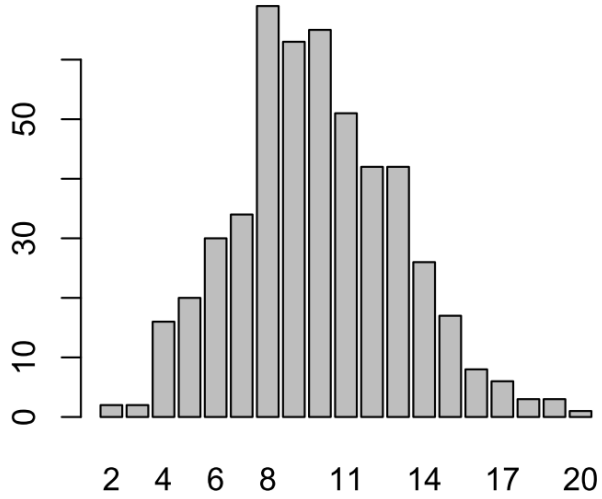
Modelos truncados en cero



λ = 1.5
P(0) = 0.22



λ = 5
P(0) = 0.007



λ = 10
P(0) = 4.5×10^{-5}

Modelos inflados en ceros

- Conteos

Propiedades:

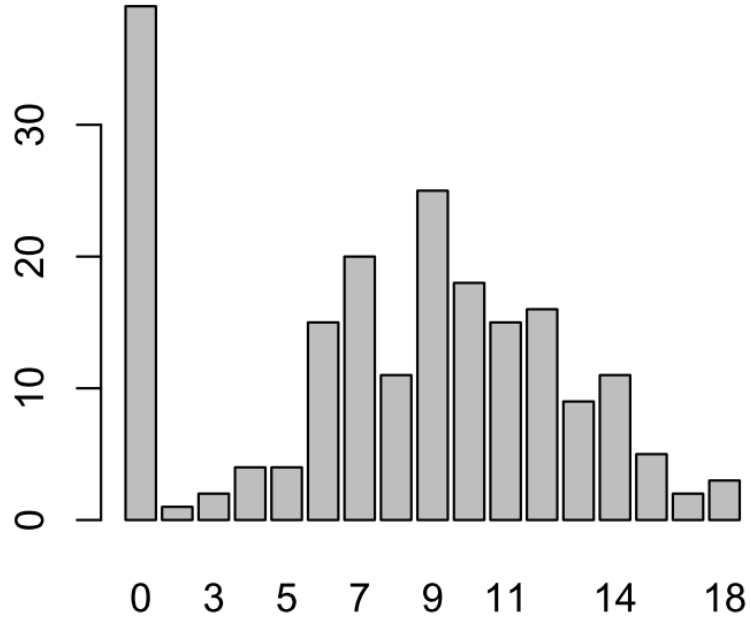
- Naturales (mínimo = 0). *Hay más ceros de lo esperado por una distribución de Poisson o binomial negativa.*
- Distribución no normal
- Distribución del error: vamos a ver…

Modelos inflados en ceros

- Problemas:

(1) Parámetros y errores estándares sesgados.

(2) Sobredispersión.



Modelos inflados en ceros

- **2 distinciones posibles:**

1. *Poisson (P) vs binomial negativo (BN)*
2. *Inflado en ceros (ZI) (= de mezcla) vs alterado en ceros (ZA) (= de 2 partes)*

- **4 modelos posibles:**

1. Poisson inflado en ceros (**ZIP**).
2. Binomial negativo inflado en ceros (**ZINB**).
3. Poisson alterado en ceros (**ZAP**).
4. Binomial negativo alterado en ceros (**ZANB**).

- Anidamiento: **ZIP** → **ZINB** y **ZAP** → **ZANB**

Modelos inflados en ceros

- En el contexto del uso de hábitat por aves en parches de bosque, Kuhnert et al. (2005) y Martin et al. (2005) distinguen los siguientes tipos de ceros:

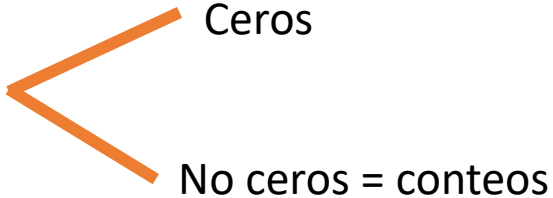
1. **Ceros estructurales** (verdaderos ceros, ceros positivos o verdaderos negativos): el ave no está presente porque el ambiente no es apto.
2. **Falsos ceros** (ceros negativos):
    - *2.a. Error de diseño ("malos ceros")*: resultado de un pobre diseño muestral o experimental. Por ejemplo, muestrear aves de pastizal en un bosque o por poco tiempo o en un área pequeña.
    - *2.b. Error del observador:* hay especies difíciles de distinguir o detectar. También es difícil detectar un ave pequeña en un bosque cerrado o en un día nublado.
    - *2.c. Error del "ave":* el ambiente es apto, pero el ave no lo utiliza.

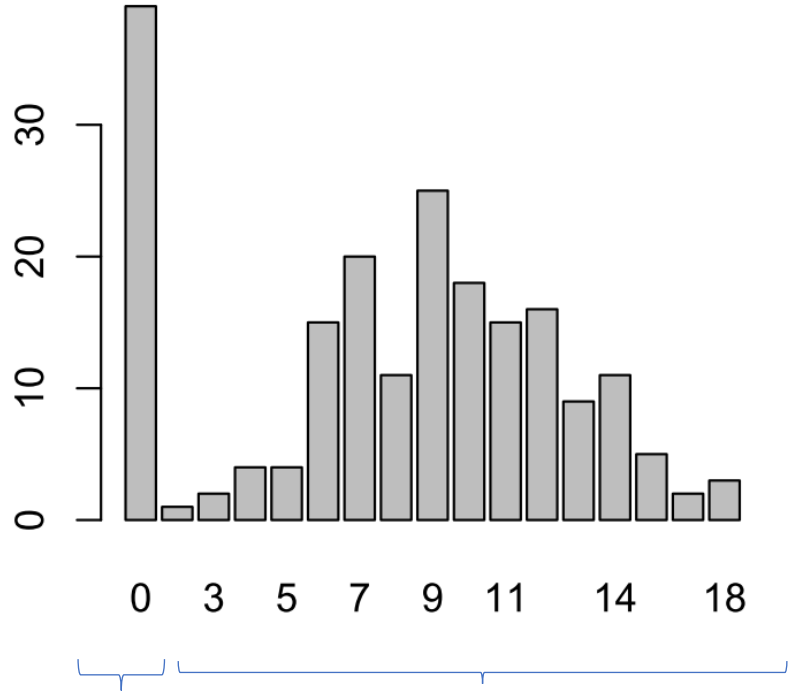
Modelos de dos partes, alterados en ceros o “valla” (ZAP y ZANB)

- Dos procesos diferentes:

1. Para modelar los ceros.
2. Para modelar los conteos.

No discrimina entre tipos de ceros.





(1) Ceros vs No ceros
(0) (1) → Binomial

(2) Conteos → Poisson o BN truncada

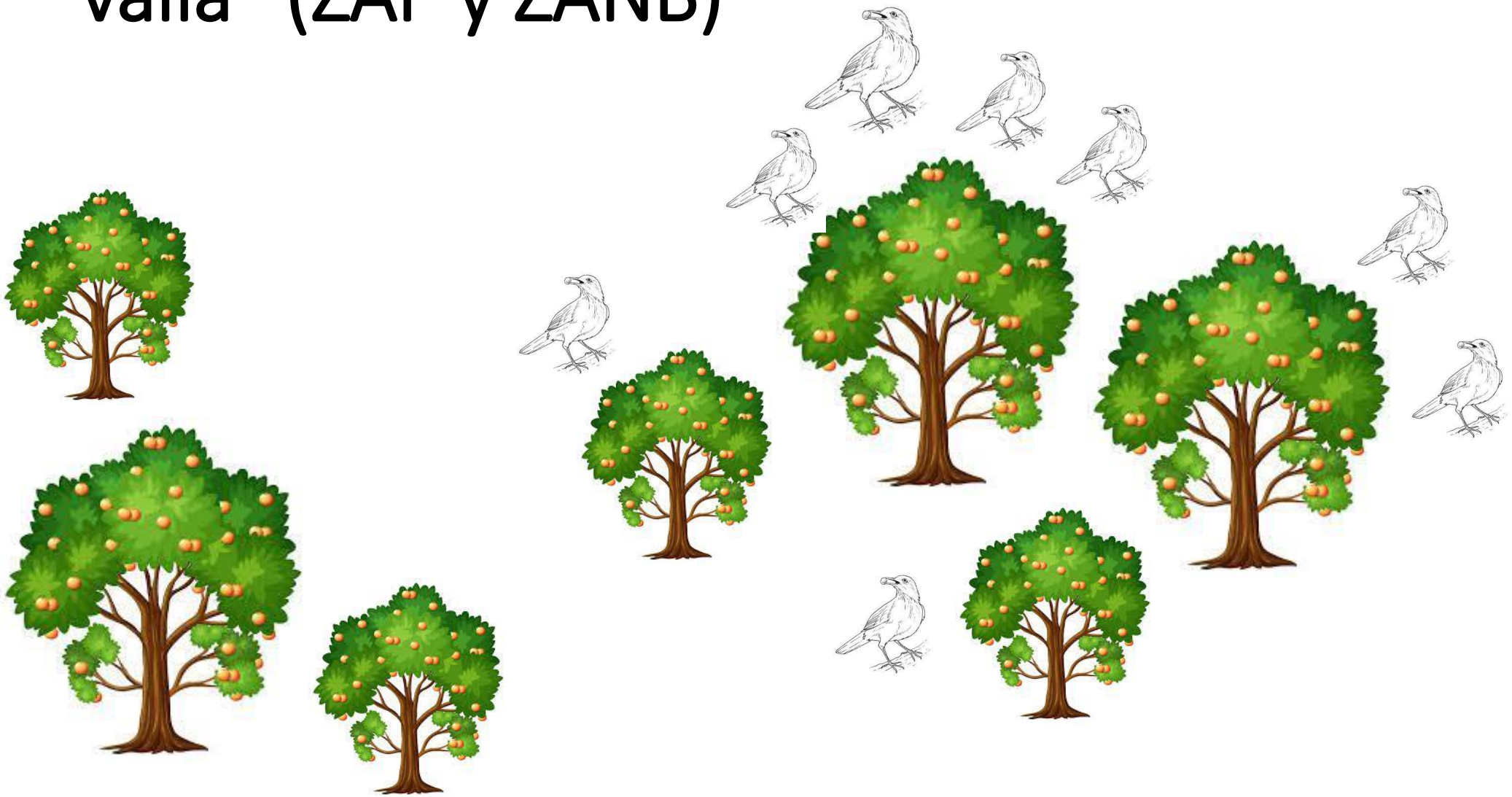
Modelos de dos partes, alterados en ceros o “valla” (ZAP y ZANB)

- 2 procesos ecológicos:

(1) Ceros observados → distribución binomial

(2) Donde no hay un cero, un segundo proceso causa los conteos → distribución Poisson o BN truncada

Modelos de dos partes, alterados en ceros o “valla” (ZAP y ZANB)

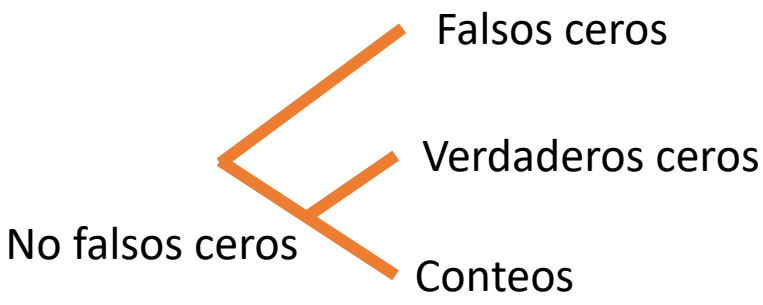


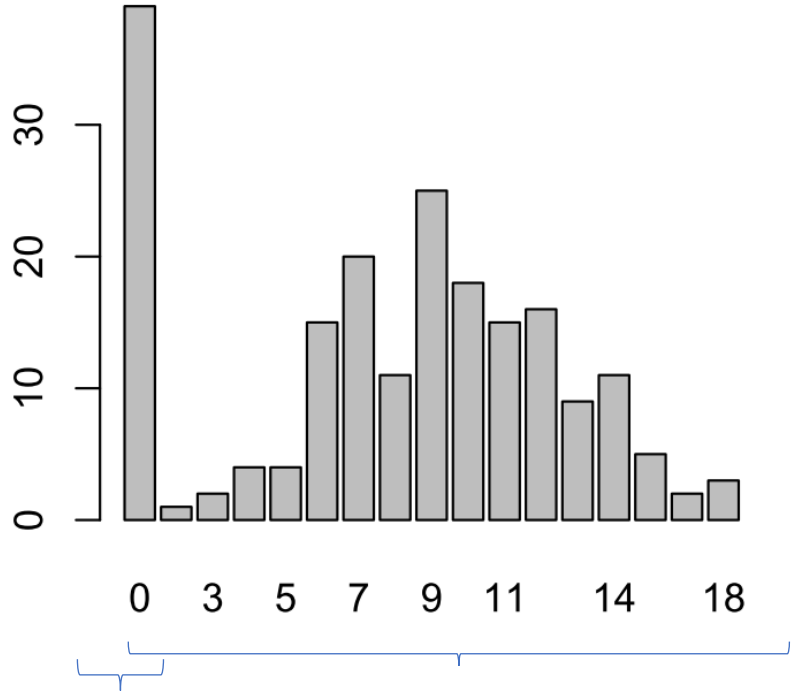
Modelos de mezcla (ZIP y ZINB)

- Dos procesos diferentes:

1. Para modelar los falsos ceros.
2. Para modelar el resto
   (verdaderos ceros + conteos)

Discrimina entre tipos de ceros.





(1) Falsos ceros vs No falsos ceros

(0) (1)

(2) Verdaderos ceros + conteos → Binomial → Poisson o BN

Modelos de mezcla (ZIP y ZINB)

Dividimos los datos en 2 conjuntos imaginarios:

1. Contiene solo (falsos) ceros (= **observaciones con masa cero**).
2. Contiene verdaderos ceros y observaciones > 0.

NO sabemos qué observaciones son verdaderos y falsos ceros

Covariables

$$\hat{p} = \frac{e^{b_0 + b_1 \text{viento}}}{1 + e^{b_0 + b_1 \text{viento}}}$$

$$\hat{y} = e^{b_2 + b_3 \text{cobertura} + b_4 \text{presas}}$$

Pueden ser diferentes para el proceso que genera los ceros y los conteos.

Sobredispersión

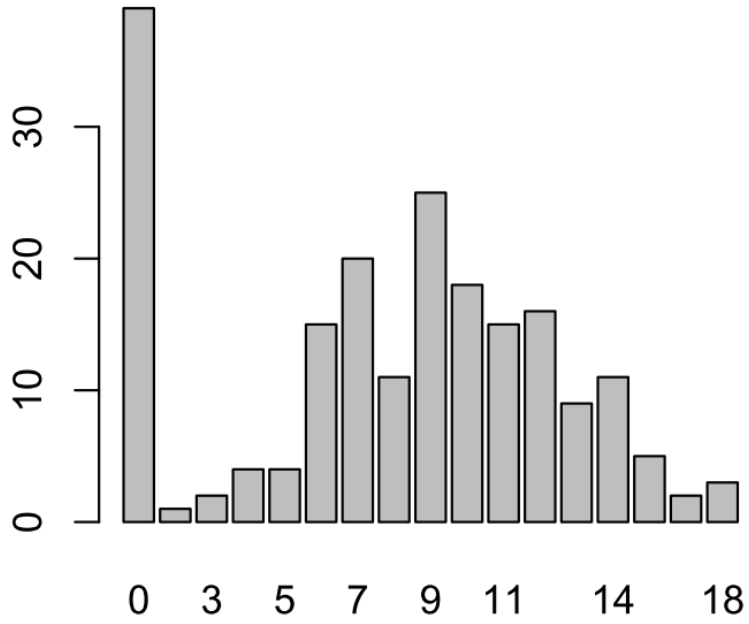
ZIP

Media = $p(1 - p)$

Varianza = $(\lambda + p\,\lambda^2)(1 - p)$

Si $p = 0$ → Varianza = λ

Si $p > 0$ → Varianza > λ



¿Cómo decidir entre GLMs Poisson, quasi-Poisson, ZIP, ZINB, ZAP y ZANB?

1. Sentido común:

¿Hay sobredispersión? → Si es pequeña → quasi-Poisson

¿Por alta variación en los conteos? → BN

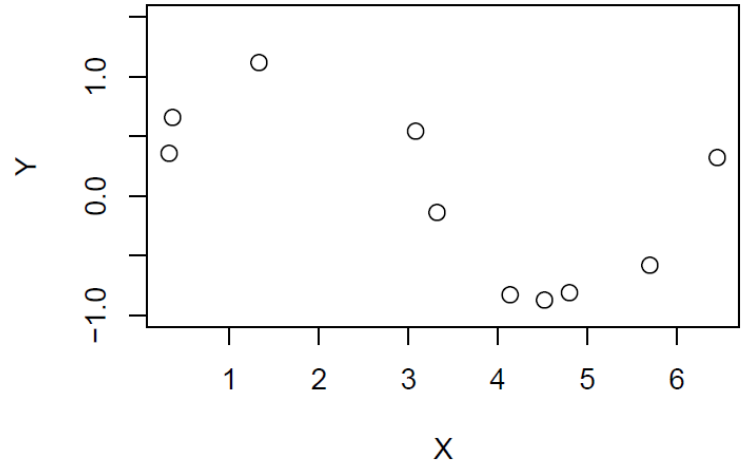
¿Por un exceso de ceros? → ZIP y ZAP

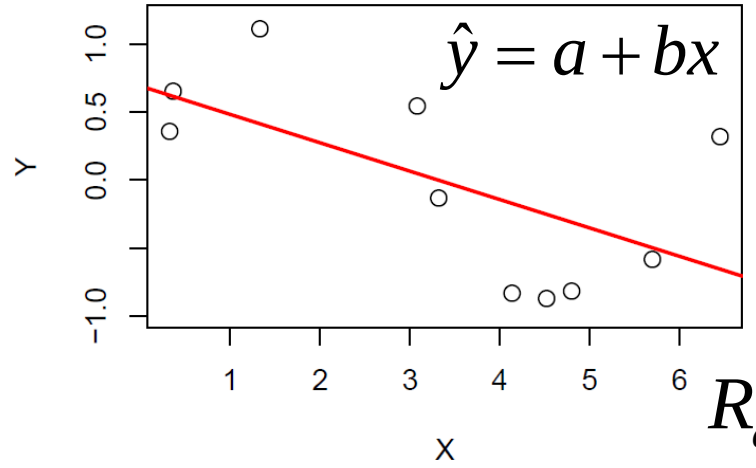
¿Variación extra en los conteos? → ZINB y ZANB

2. Tests de razón de verosimilitud → Entre ZIP y ZINB, entre ZAP y ZANB

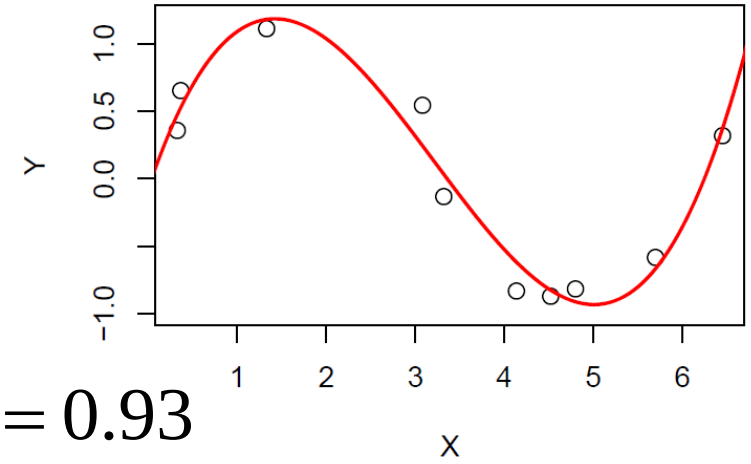
3. Criterios de información.

¿Qué es un buen modelo?



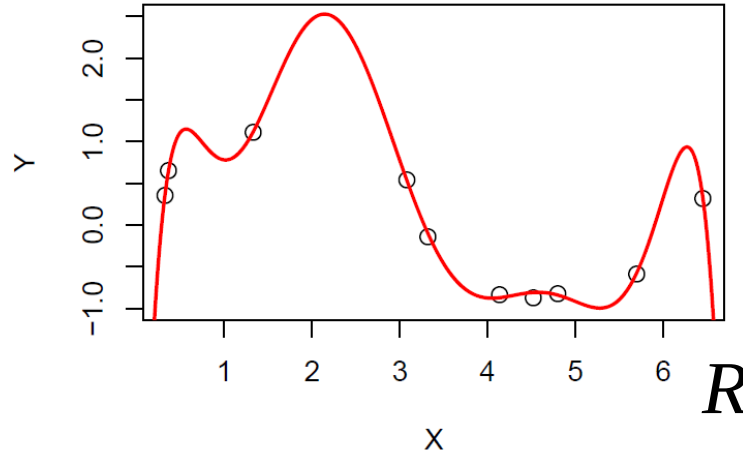


$R_a^2 = 0.31$



$R_a^2 = 0.93$

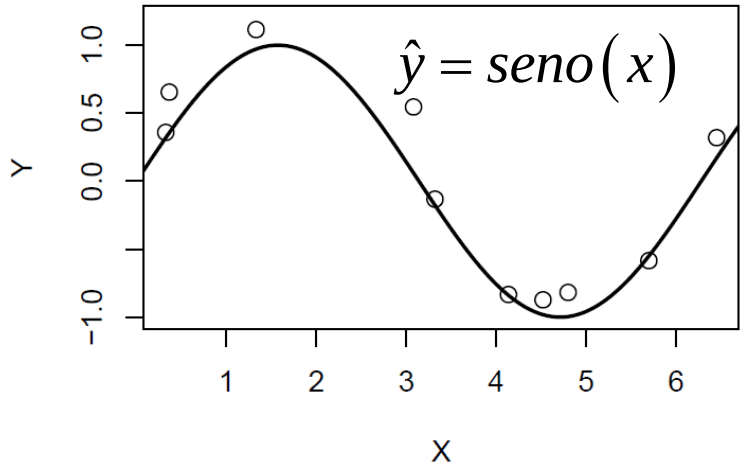
$$\hat{y} = a + b_1 x + b_2 x^2 + b_3 x^3$$

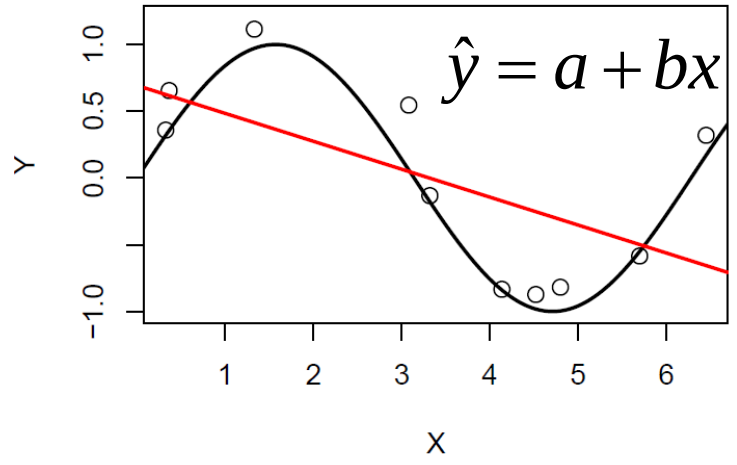


$R_a^2 = 0.98$

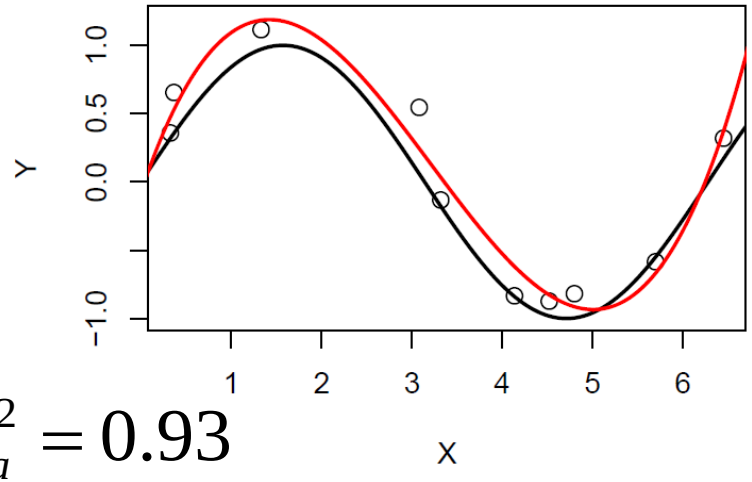
$$\hat{y} = a + b_1 x + b_2 x^2 + b_3 x^3 + b_4 x^4 + b_5 x^5 + b_6 x^6 + b_7 x^7 + b_8 x^8$$

¿Qué es un buen modelo?



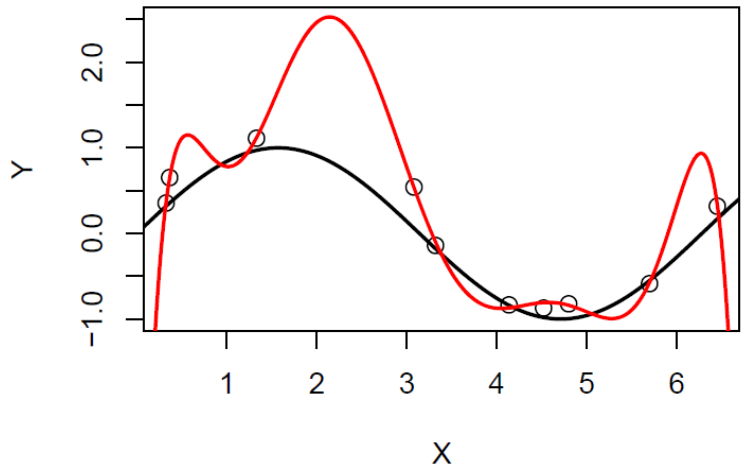


$R_a^2 = 0.31$



$R_a^2 = 0.93$

$\hat{y} = a + b_1 x + b_2 x^2 + b_3 x^3$



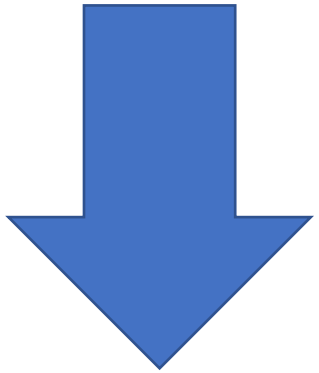
$R_a^2 = 0.98$

$\hat{y} = a + b_1 X + b_2 X^2 + b_3 X^3 + b_4 X^4 +$

$b_5 X^5 + b_6 X^6 + b_7 X^7 + b_8 X^8$

Inferencia multimodelo

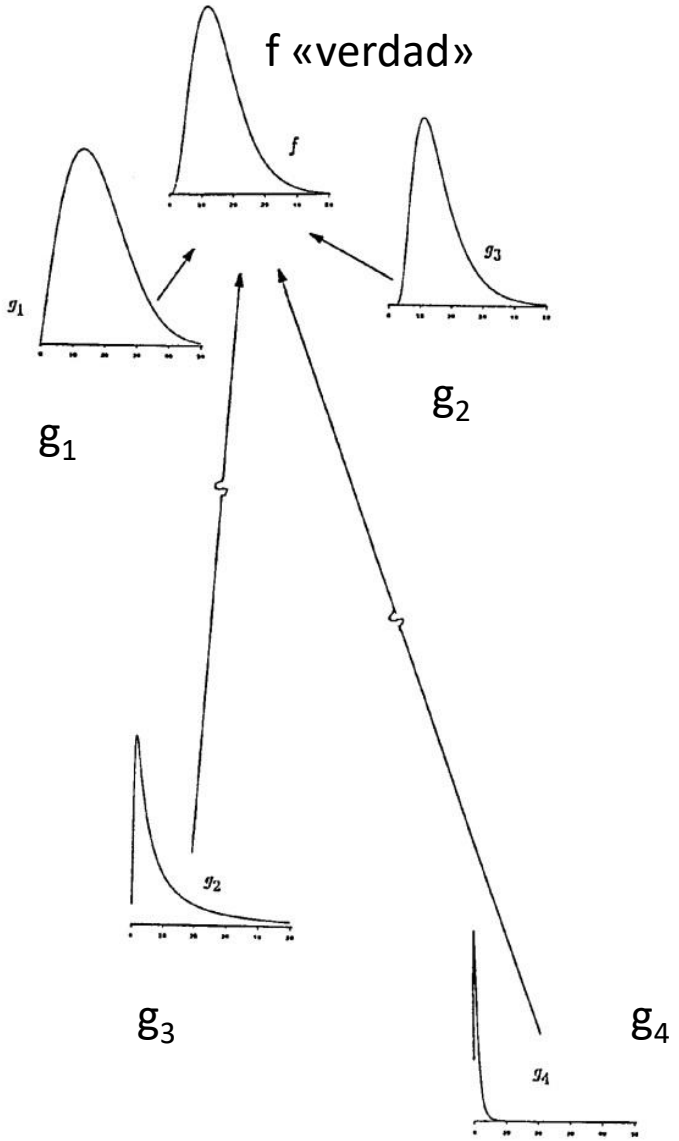
¿Qué es un “buen” modelo?



Modelo que explica la mayor parte de la variación con el menor número de parámetros (principio de parsimonia).

“*Todos los modelos son erróneos, pero algunos son útiles*” (George Box)

Información de Kullback-Leibler



Información perdida cuando se utiliza *g* para aproximar *f*

$g_4 > g_3 > g_1 > g_2$

Pero… no se conoce el modelo verdadero que dio origen a los datos.

Criterio de Información de Akaike

- Akaike (1973) propuso el uso de la información K-L como una base fundamental de la selección de modelos.
- Se puede utilizar la información K-L relativa en lugar de la absoluta.

$$\mathrm{AIC} = -2\log L + 2K$$

¿Cuánta más información se pierde por g_1 que por g_2?

Criterio de información de Akaike

$$\mathrm{AIC} = -2\log L + 2K$$

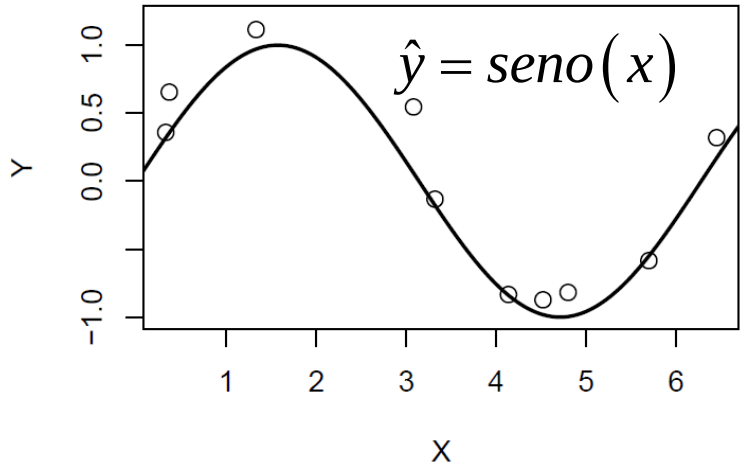
Ajuste

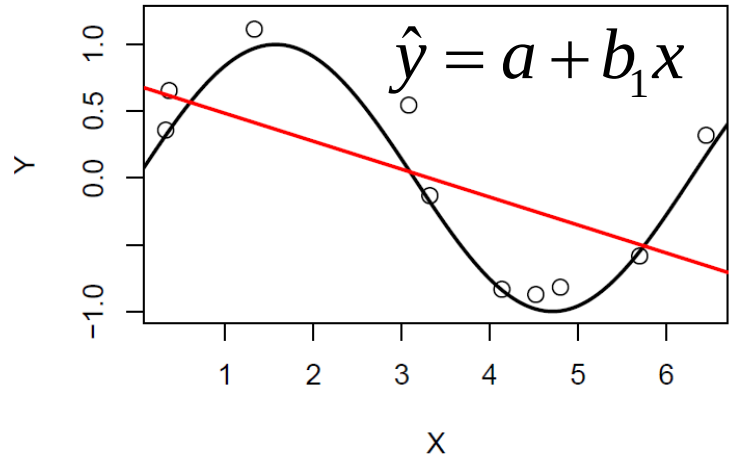
Cantidad de parámetros

- NO habla de hipótesis nulas
- Calidad de un modelo, relativa a otros modelos
- Si todos los modelos son malos, aún así el AIC encontrará el mejor modelo

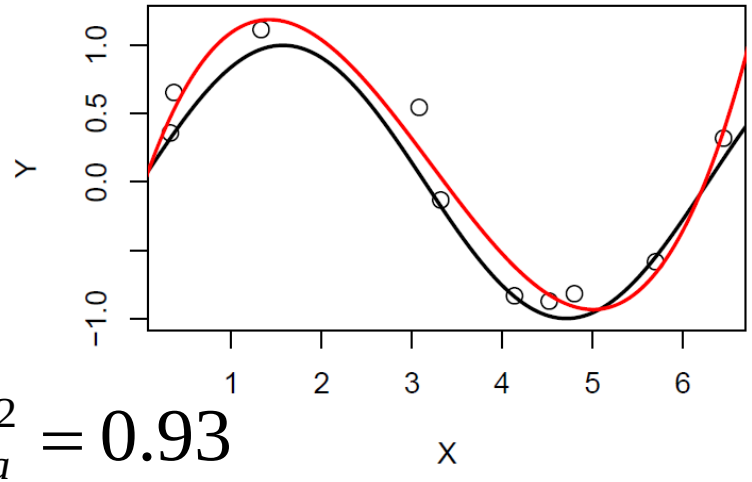
$$\mathrm{AICc} = \mathrm{AIC} + \frac{2K\left(K+1\right)}{n-K-1}$$

¿Qué es un buen modelo?



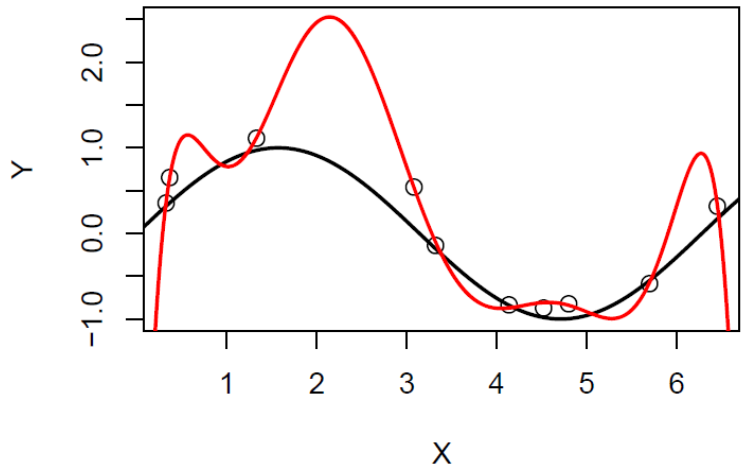


$R_a^2 = 0.31$



$R_a^2 = 0.93$

$$\hat{y} = a + b_1 x + b_2 x^2 + b_3 x^3$$



$R_a^2 = 0.98$

$$\hat{y} = a + b_1 x + b_2 x^2 + b_3 x^3 + b_4 x^4 + b_5 x^5 + b_6 x^6 + b_7 x^7 + b_8 x^8$$

¿Cómo comparamos modelos?

Hipótesis biológicas → *n* modelos

y: abundancia de una sp de insecto
x1: cobertura vegetal
x2: densidad de plantas
x3: disponibilidad de presas

Modelo	Hipótesis
y ~ 1	Nula
y ~ x1 + x2	El ambiente determina la abundancia, y no la disponibilidad de presas
y ~ x1 + x2 + x3	Tanto el ambiente como la disponibilidad de presas determinan la abundancia
y ~ x3	La disponibilidad de presas determina principalmente la abundancia, y no el ambiente

¿Cómo comparamos modelos?

Modelo	Hipótesis	AIC	ΔAIC	w
y ~ x1 + x2	El ambiente determina la abundancia, y no la disponibilidad de presas	1028	0.0	0.62
y ~ x1 + x2 + x3	Tanto el ambiente como la disponibilidad de presas determinan la abundancia	1029	1.0	0.38
y ~ 1	Ninguna de las variables explica la abundancia	1042	14.0	<0.01
y ~ x3	La disponibilidad de presas determina principalmente la abundancia, y no el ambiente	1049	21.0	<0.01

$$w_i = \frac{e^{-\frac{1}{2}\Delta_i}}{\sum e^{-\frac{1}{2}\Delta_i}} = c\frac{1}{e^{0.5\Delta_i}}$$

Probabilidad de ser el mejor modelo entre los modelos candidatos

¿Podemos comparar todos los modelos posibles?

- NO hacer “dragado” de modelos

“ “Dejar a la computadora que lo descubra” es una estrategia pobre y generalmente refleja el hecho de que el investigador no se molestó en pensar claramente acerca del problema de interés y de su escenario científico” (Burnham y Anderson 2002)

Promediado de modelos

¿Cuándo promediar? w < 0.9 (Grueber et al. 2011)

Modelo	x1	x2	x3
1	0.2		-0.6
2		0.3	-0.4

¿Cómo promediar?

2 formas:

1. “Modelo condicional” o método “natural” → y = 0.2x1 + 0.3x2 – 0.5x3
2. “Modelo completo” o método “cero” → y = 0.1x1 + 0.15x2 – 0.5x3

Promediado de modelos

Nakagawa y Freckleton (2011)

- Método “cero” → determinar factores con la mayor influencia sobre la respuesta.
- Método “natural” → determinar el efecto de una variable en particular (posiblemente con un efecto débil) sobre la variable respuesta.

Resumen sobre los principios de selección basada en AIC

- 1. AIC es una medida relativa de parsimonia. ↓AIC, mejor
- 2. Podemos comparar modelos anidados o no anidados.
- 3. No comparar demasiados modelos. Se puede encontrar un mejor modelo por azar que sea malo. Basarse en el sentido biológico.
- 4. Es posible tener varios mejores modelos que pueden o no promediarse.
- 5. Modelos con bajo tamaño muestral ($n/k < 40$) utilizar AICc.
- 6. El mejor modelo puede tener poca capacidad explicatoria.
- 7. Es el más utilizado en ecología, pero no el único (BIC, DIC, WAIC).


P>0.05
AIC