

# Análisis Estadístico Lineal Avanzado en R

Curso Colaborativo IIAP - UNAMAD

Irwing S. Saldaña

Instituto de Ciencias Antonio Brack

Departamento de Ecoinformática y Biogeografía

Perú, 2021



# Blgo. Irwing S. Saldaña

## Instructor

Dpto. de Ecoinformática y Biogeografía,  
Instituto de Ciencias Antonio Brack, Perú

[Website](#) | [ResearchGate](#) | [Linkedin](#) | [R Latam Blog](#) | [Github](#)



# George Box

"...en esencia , todos los modelos están equivocados, pero algunos son útiles."



# Principio de Parsimonia

"En igualdad de condiciones, la explicación más sencilla suele ser la correcta"



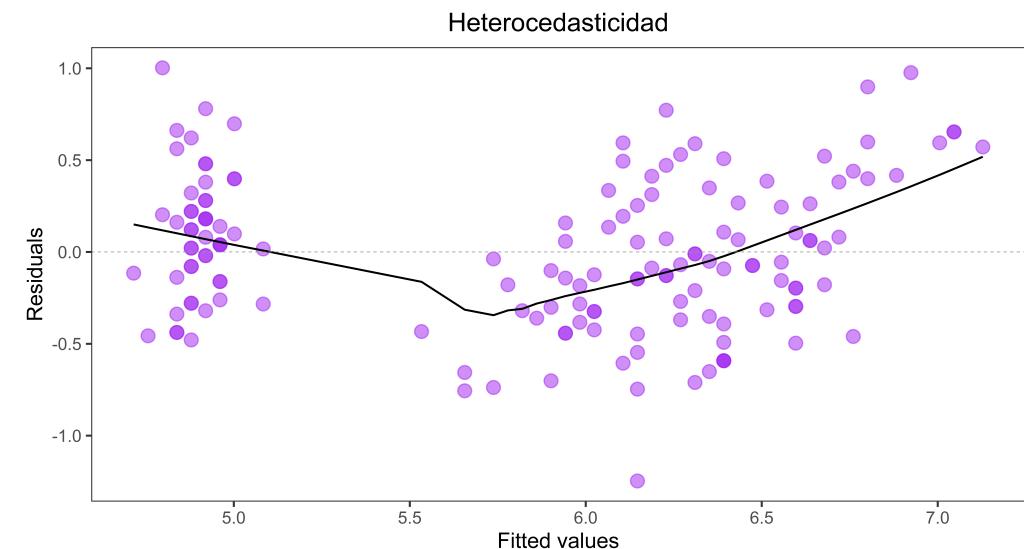
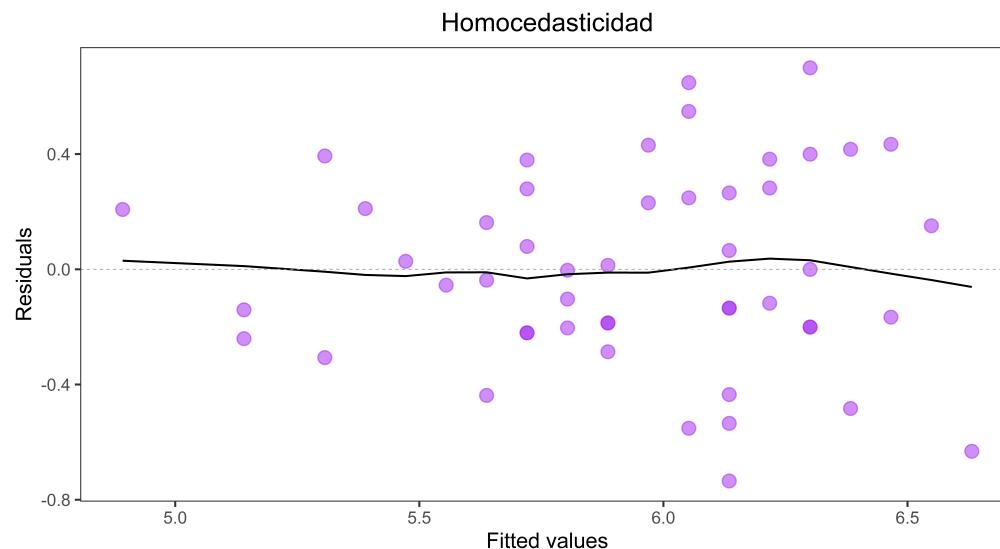
# Introducción a los Modelos Lineales de Efectos Mixtos (LMM)

[ Ajustando temas de dependencia en los modelos lineales ]



# Homocedasticidad en los modelos lineales

Cuando hay **homocedasticidad**, la varianza de la variable respuesta (Y) es constante a lo largo de la población: no depende del valor de X.



Lo opuesto es **Heterocedasticidad**, y ocurre comúnmente cuando la data no es independiente, evidenciando un efecto de **agrupamiento** en la colecta de la data.



# Enfoques de análisis

## 1. Regresiones lineales

(grupos separados)  
Cada A. en cada B.

2 parámetros x 3 grupos A x 6  
sitios B = 36 estimadores  
(coeficientes) a los que  
interpretar.

Errores diferentes en cada  
modelo. Imposible tomar en  
cuenta todo a la vez.

## 2. Regresiones lineales

Con toda la base de datos sin tomar en  
cuenta que fueron tomadas en  
diferentes A y B.

2 parámetros, entre 3 y 4,  
(coeficientes) para interpretar.

Demasiado ruido en los modelos  
debido a que la autocorrelación  
de las muestras dentro de cada A  
y B influenciarían demasiado el  
resultado.

## 3. Modelo Lineal de Efectos Mixtos

2 parámetros, pocos  
estimadores, entre 3 y 4, para  
interpretar.

Perfecto para este tipo de  
situaciones. Toma en cuenta la  
variabilidad de todos las  
agrupaciones, reduciendo el  
ruido en los resultados.

> Blgo. Irwing S Saldaña [Programa de Certificación Especializado Data Science: Estadística y Análisis de Datos en R]



# Nomenclatura de Efectos Mixtos

Se llaman así debido a que involucran dos tipos de variables en el componente sistemático de las fórmulas:

## Efectos fijos:

- Es una variable categórica o numérica.
- Variables de las cuales estamos interesados en obtener conclusiones.
  - Si es categórica, realiza mediciones en todos los posibles niveles que esa variable pueda tener según mi estudio.
  - Si es un factor, estamos interesados en obtener información sobre todos los niveles ( e.g ., tratamientos) para los que se recopilaron los datos.

## Efectos aleatorios:

- Es una variable categórica (variables de agrupamiento).
- Variables de las cuales NO estamos interesados en obtener conclusiones, pero que generan variabilidad.



# Topología de modelos `lmer()` en R

## Modelos clásicos de solo efectos fijos (LM):

- En los modelos lineales clásicos  $y = \beta_0 + \sum(\beta_i x_i)$  se considera a  $\beta_0$  y  $\beta_i$  como efectos fijos.

```
# Ejemplo en R con dos variables explicativas (Efectos fijos)
lm(Var_respuesta ~ Efec_fijo + Efec_fijo, data=DF)
```

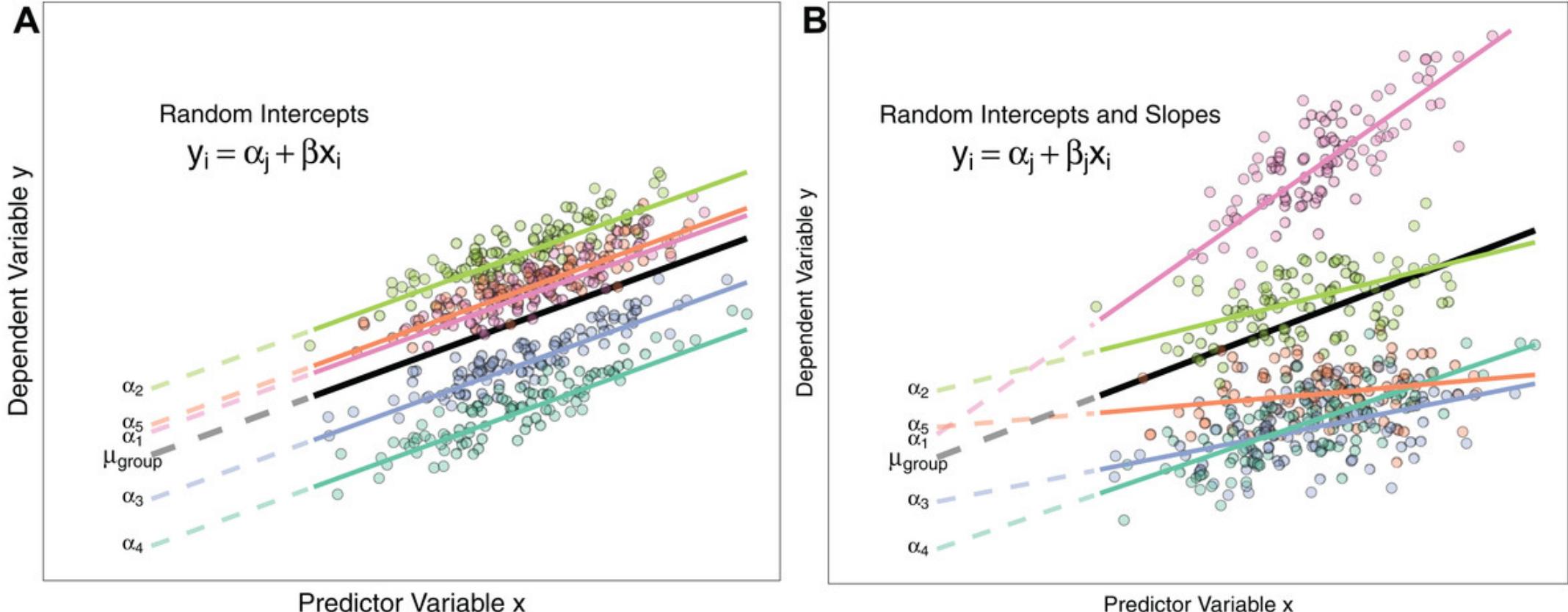
## Modelos con efectos mixtos (LMM):

- Son variables que generan variabilidad en la varianza del modelo por el efecto de procesos de agrupamiento, anidamiento o cruzamiento en la toma de datos.
  - En el modelo se incluyen como  $y = \beta_0 + \sum(\beta_i x_i) + Z_u$

```
# Ejemplo en R con dos variables explicativas (Ambos tipos de efectos)
lm(Var_respuesta ~ Efec_fijo + (1|Efec_Aleatorio), data=DF)
```



# Modelos Lineales de Efectos Mixtos



Fuent: rinterceptyslope.jpg

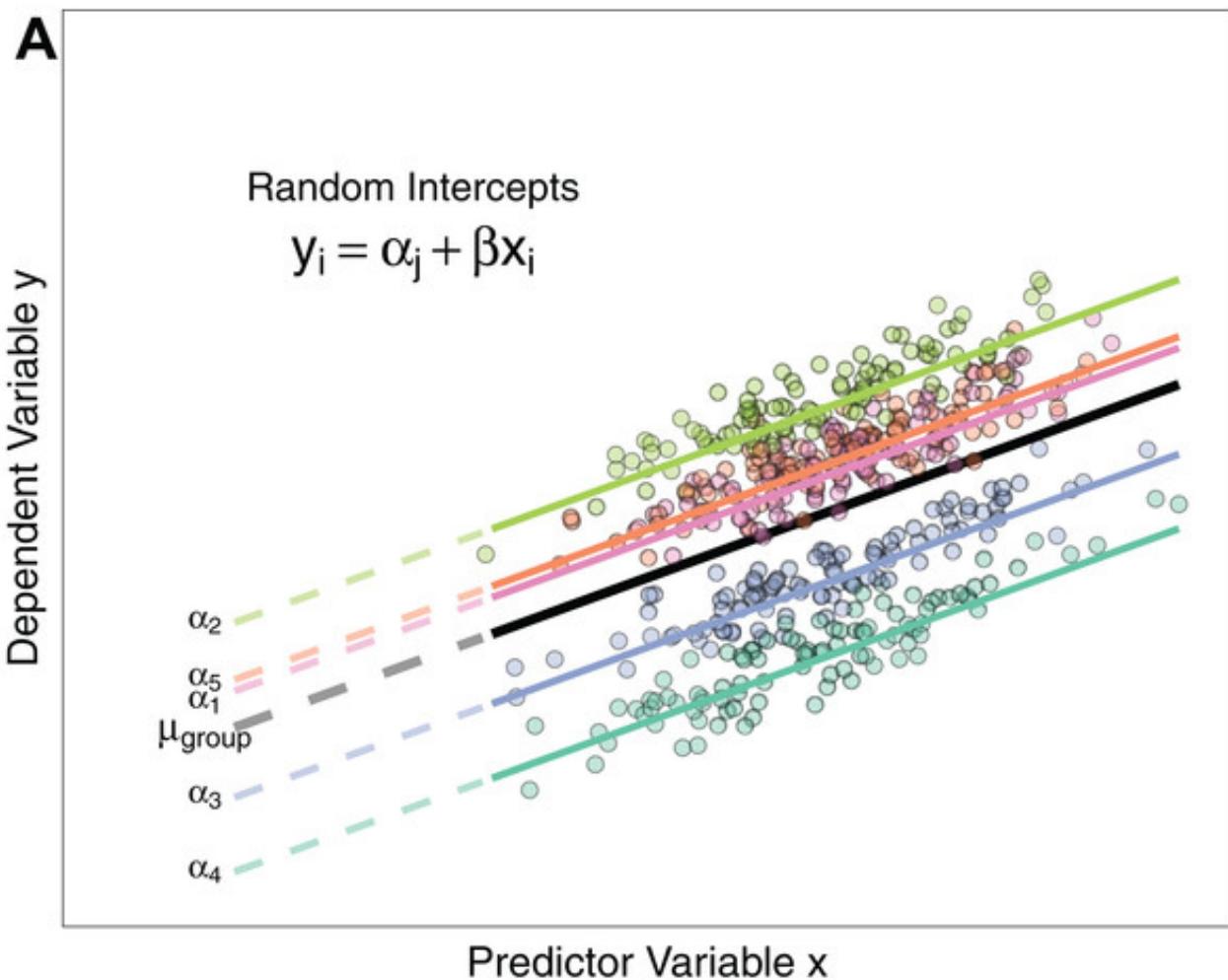


# Modelo LMM con Interceptos Aleatorios

```
lmer(y ~ x + (1|v.aleatoria), data=DF)
```

- Se permite que el intercepto varíe para cada grupo.
- Se asume que la variable explicativa tiene la misma tendencia de efecto para cada grupo, pero el promedio esperado difiere según la observación pertenezca a uno u otro grupo.

$$y = \beta_0 + v_0 + \beta_1 x_1 + \epsilon$$

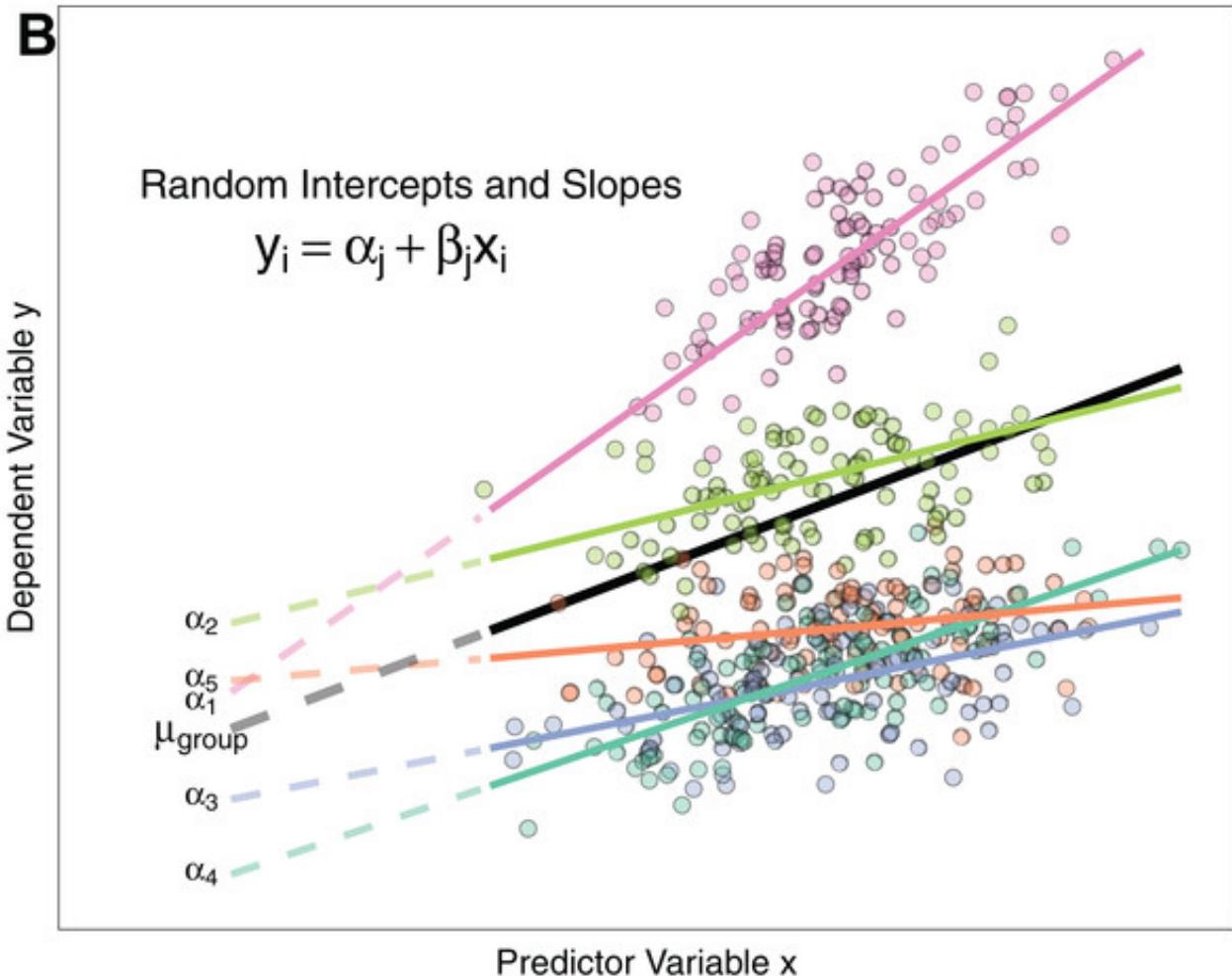


# Modelo LMM con Interceptos y Pendientes Aleatorias

```
lmer(y ~ x + (x|v.aleatoria), data=DF)
```

- Se usa cuando se conoce que la variable explicativa tiene efectos diferentes en los diferentes grupos evaluados.
- Esto genera que cada uno tenga una pendiente diferente.

$$y = \beta_0 + v_0 + (\beta_1 + v_1)x_1 + \epsilon$$



# Desarrollemos la práctica

60 : 00



# Introducción a los Modelos Lineales Generalizados (GLM)

[ Expandiendo los límites de un modelo lineal ]



# Modelos Generalizados Lineales

Recordemos, en los modelos lineales **LM** asumimos que:

$$Y|(X_1 = x_1, \dots, X_p = x_p) \sim N(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p, \sigma^2)$$

Por lo tanto

$$\mathbb{E}[Y|X_1 = x_1, \dots, X_p = x_p] = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p, \sigma^2 = \eta.$$

- La condición necesaria para satisfacer que el error sea normal es que  $Y$  sea continua.

**Los GLM buscan ampliar las limitantes de los LM para incorporar análisis variables respuesta principalmente discretas (binarias o conteos).**



# Definición Matemática de un GLM

Para definir un **GLM** necesitamos dos componentes

- **Componente respuesta  $Y$  (variable respuesta, dependiente):** Las observaciones de  $y$  son independientes y provienen de una familia de dispersión exponencial (FDE)

$$Y|(X_1 = x_1, \dots, X_p = x_p) \sim FED(\mu, \phi)$$

- **Componente sistemático  $\eta$  (variable(s) explicativa(s), independiente(s)):**  Es la parte de la ecuación definida por las variables explicativas, los coeficientes, la pendiente y el error. Sigue siendo un componente lineal.

$$\eta = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \epsilon$$

Por tanto, el modelo GLM se define como función de la variable  $y$  predicha  $g(\mu)$

$$g(\mu) = \eta = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \epsilon$$



# Función de enlace

La función  $g(\mu)$  es conocida como **función de enlace (link function)**. Esta es reversible, por lo que podemos transformar los coeficientes resultantes para calcular el promedio de calculado para  $Y$  como  $g^{-1}(\eta) = \mu$

Familia	Función de enlace	Función en R	Argumento family/link
Gausiana	Identity	glm()	gaussian(link='identity')
Binomial	Logit, Probit, cloglog	glm()	binomial(link='logit')
Poisson	Log, identity, sqrt	glm()	poisson(link='log')
Gamma	Inverse, identity, Log	glm()	Gamma(link='inverse')
Inversa Gausiana	Inverse, identity, Log	glm()	inverse.gaussian(link = '1/mu^2')
Quasibinomial	Logit, Probit, cloglog	glm()	quasibinomial(link = 'logit')
Quasipoisson	Log, identity, sqrt	glm()	quasipoisson(link = 'log')
Binomial Negativa	Log, identity, sqrt	glm.nb()	link = 'log'



Las funciones canónicas (**definidas por defecto**) de las familias de distribución de probabilidades se describen a continuación. Es importante también saber cuál es el valor y calculado en el Rango de cada función.

Distribución	Rango de $Y$	Notación de distribución	Función de enlace canónica $g(\mu)$	Valor esperado $\mathbb{E}(Y)$	Param. Varianza $\phi$
Gaussiana	$\mathbb{R}$	$\mathcal{N}(\mu, \sigma^2)$	$\mu$	$\eta$	$\sigma^2$
Bernoulli	$0, 1$	$\text{Ber}(p)$	$\text{logit}(\mu)$	$\text{logistic}(\eta)$	$1$
Binomial	$0, \dots, N$	$\text{B}(N, p)$	$\log\left(\frac{\mu}{N-\mu}\right)$	$N \cdot \text{logistic}(\eta)$	$1$
Poisson	$0, 1, \dots$	$\text{Pois}(\lambda)$	$\log(\mu)$	$e^\eta$	$1$
Gamma	$(0, \infty)$	$\Gamma(a, \nu)$	$-\frac{1}{\mu}$	$-\frac{1}{\eta}$	$\frac{1}{v}$



# Topología de los función `glm()` en R

```
# Estructura generalizada  
glm(formula, data = DF, family = `familia(link = "función.de.enlace")`)  
  
# Regresión logística (ejemplo)  
glm(y ~ x1 + x2, data = DF, family = binomial(link = "logit"))  
  
# Regresión de Poisson (ejemplo)  
glm(y ~ x1 + x2, data = DF, family = poisson(link = "log"))
```

- Las **fórmulas** definen el enfrentamiento de las variables explicativas  $X$  y la variable respuesta  $Y$ .



# Regresiones GLM

- Regresión Logística (Bernoulli, Binomial)
- Regresión de Poisson
- Regresión Binomial Negativa
- Regresión de Hurdle
- Regresión Zeroinflado
- Regresión Beta
- Regresión Gaussiana (  $\equiv$  Regresión Lineal Clásica)
- Regresión Gamma
- ...

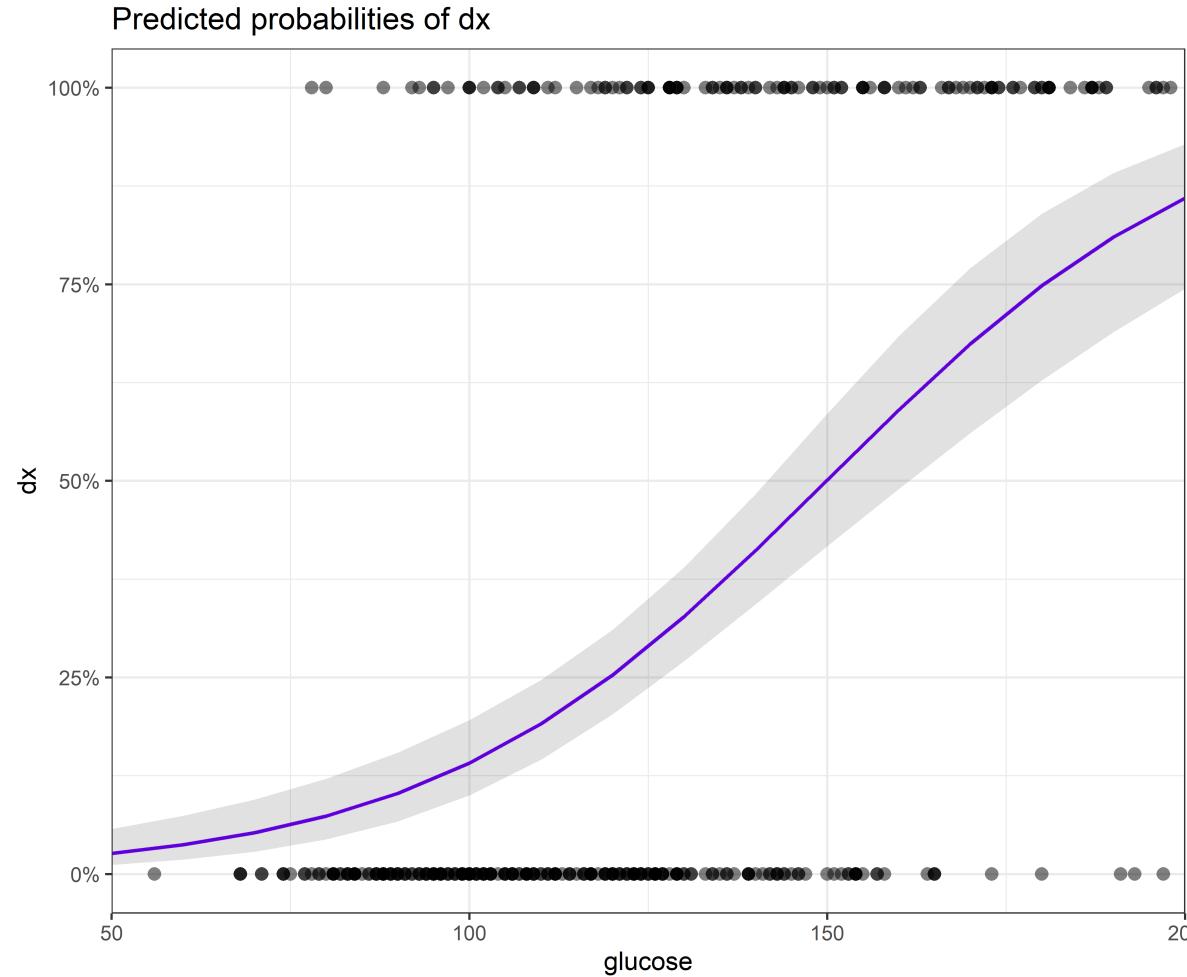


# Regresión Logística

[ Modelamiento de respuestas discretas binarias ]



# ¿Cómo luce una regresión logística?



# Regresión Logistica

- La variable respuesta  $Y$  es binaria (Distribución de Probabilidades de Bernoulli)

$$Y = \begin{cases} 1, & \text{(evento) con probabilidad } p, \\ 0, & \text{(no evento) con probabilidad } 1 - p, \end{cases}$$

- Si asumimos que

$$\mathbb{E}[Y|X_1 = x_1, \dots, X_p = x_p] = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p, \sigma^2 = \eta$$

Estaríamos aceptando predicciones negativas y positivas fuera del rango  $[0, 1]$ :

- Para evitar tal error, se utiliza una **función de enlace**  $g()$  que **projete**  $Y$  sobre los números  $\mathbb{R} (-\infty, +\infty)$ , y su operación inversa  $g^{-1}()$  para **devolver** las predicciones del modelo al rango  $[0, 1]$ .



# Concepto de Probabilidad

- **Probabilidad:** es la probabilidad de que un evento suceda.

$$p = 0.8$$

Imagina que ese  $p$  significa un 80% de probabilidad de que llueva en un día nublado.

- Para **interpretar los resultados** de la regresión logística, debemos introducir la terminología adecuada para ello: los **odds**.



# Concepto de Odds

- **Odds:** es la **posibilidad de éxito** de un evento. Está definido como la probabilidad de que el evento ocurra dividida por la probabilidad de que este mismo evento no ocurra.

$$\text{odds}(Y) := \frac{p}{1 - p} = \frac{\mathbb{P}[Y = 1]}{\mathbb{P}[Y = 0]}$$

Si existe un 80% de probabilidad de que llueva en un día nublado y 20% de que no las posibilidades de que llueva en un día nublado son:

$$odds = \frac{0.8}{1 - 0.8} = \frac{0.8}{0.2} = 4$$

La posibilidad de que llueva en un día nublado es 4 veces la posibilidad de que no llueva.



# Concepto de Odds ratio

- **Odds ratio:** es el **ratio o división de dos odds** asociados. Está definido como la probabilidad de que el evento ocurra dividida por la probabilidad de que este mismo evento no ocurra.

$$\text{odds ratio} = \frac{\text{odds}_1}{\text{odds}_2} = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}$$

Si consideramos que la probabilidad de que llueva  $p_2$  en un día soleado es de 10%

$$\text{odds}_2 = \frac{p_2}{1 - p_2} = \frac{0.1}{1 - 0.1} = \frac{0.1}{0.9} = 0.111\dots$$

En consecuencia

$$\text{odds ratio} = \frac{\text{odds}_1}{\text{odds}_2} = \frac{4}{0.111} = 36$$

El ratio de las posibilidades de que llueva en un día nublado es 36 veces la posibilidad de llueva en un día soleado



# Por qué Odds y Log odds

$$odds = \frac{p}{1 - p}$$

$$\log(odds) = \log\left(\frac{p}{1 - p}\right)$$

$$\log(odds) = \log\left(\frac{p}{1 - p}\right)$$

si  $p = 0$ , entonces

$$\log(odds) = \log\left(\frac{0}{1 - 0}\right)$$

$$\log(odds) = \log\left(\frac{0}{1}\right)$$

$$\log(odds) = \log(0) - \log(1)$$

$$\log(odds) = -\infty - 0$$

$$\log(odds) = -\infty$$

$$\log(odds) = \log\left(\frac{p}{1 - p}\right)$$

si  $p = 1$ , entonces

$$\log(odds) = \log\left(\frac{1}{1 - 1}\right)$$

$$\log(odds) = \log\left(\frac{1}{0}\right)$$

$$\log(odds) = \log(1) - \log(0)$$

$$\log(odds) = 0 - -\infty$$

$$\log(odds) = +\infty$$



# Función Logit

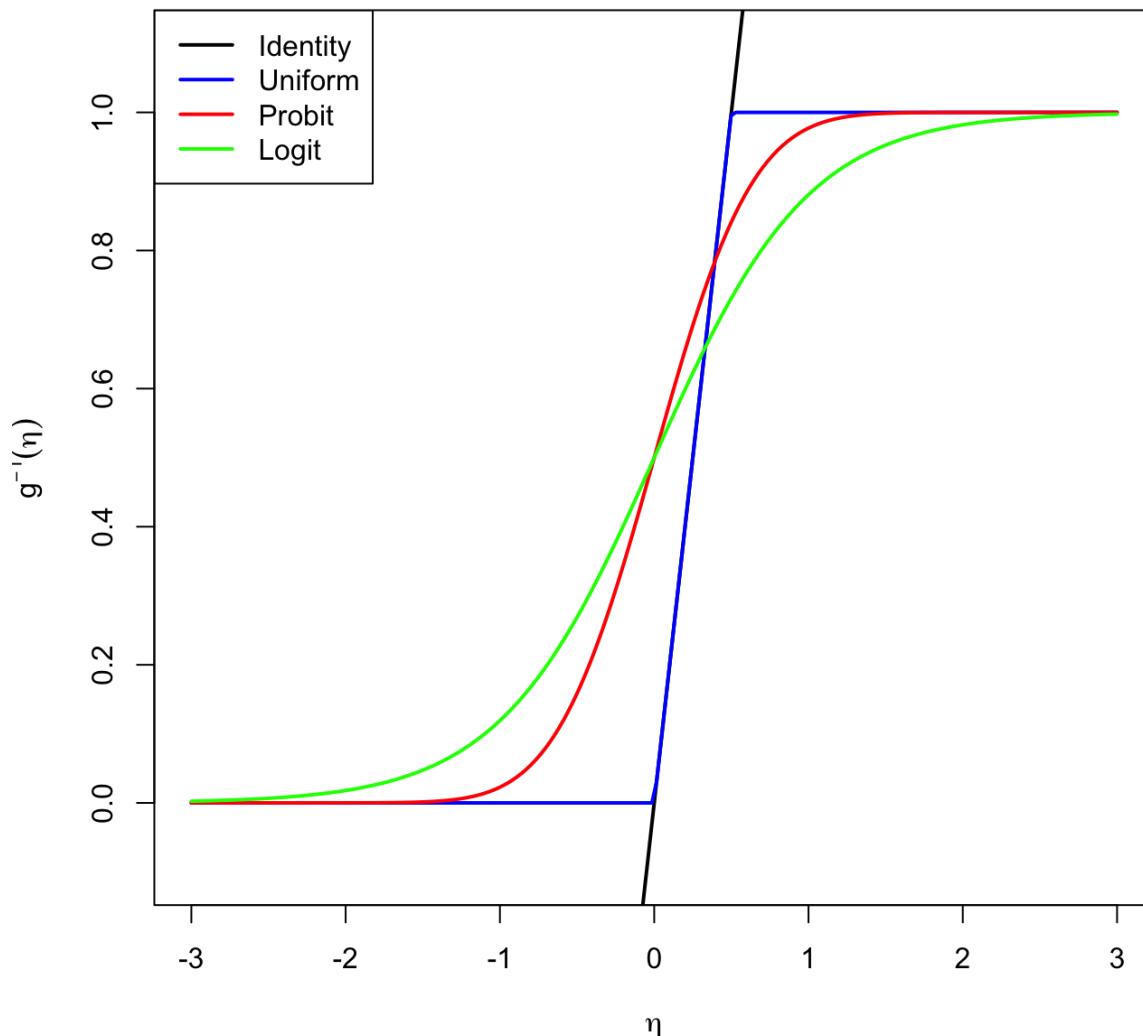
- Proyecta valores entre  $[0, 1]$  en los  $\mathbb{R}$

$$\text{logit}(\mathbb{E}[Y|X_1 = x_1, \dots, X_p = x_p]) = \eta$$

- Siendo que,

$$\eta = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p.$$

- Para ser más precisos en la interpretación de los coeficientes, necesitamos introducir las posibilidades (**los odds**).



# Intepretar los coeficientes (R. logística)

- $\beta_0$  : son los log-odds cuando  $X_1 = \dots = X_p = 0$
- $\beta_j$  es el incremento de los log-odds por cada unidad de incremento de la variable  $X_j$  cuando las demás variables explicativas se mantienen constantes en cero (no cambian).

Como estos valores no son sencillos de interpretar, se deben **transformar**:

- $e^{\beta_0}$  : son los odds cuando  $X_1 = \dots = X_j = 0$
- $e^{\beta_j}$  : es el incremento multiplicativo de los odds por cada unidad de incremento de la variable  $X_j$  cuando las demás variables explicativas se mantienen constantes en cero (no cambian).

Interpretar los resultados a primera vista nos puede dar una idea del efecto de  $X_j$  sobre  $\mathbb{E}[Y]$

- Si:  $\beta_j > 0 \rightarrow e^{\beta_j} > 1 \rightarrow$  incremento de la posibilidad
- Si:  $\beta_j < 0 \rightarrow e^{\beta_j} < 1 \rightarrow$  disminución de la posibilidad
- Si:  $\beta_j = 0 \rightarrow e^{\beta_j} = 1 \rightarrow$  efecto nulo



Si tengo **odds**, puedo convertirlo en probabilidad usando la fórmula:

$$\text{odds} = e^{\beta_j}$$

$$e^{\beta_j} = \frac{p}{1 - p}$$

$$e^{\beta_j} * (1 - p) = p$$

$$e^{\beta_j} - e^{\beta_j} * p = p$$

$$e^{\beta_j} = p + e^{\beta_j} * p$$

$$e^{\beta_j} = p * (1 + e^{\beta_j})$$

$$p = \frac{e^{\beta_j}}{(1 + e^{\beta_j})}$$



Si tengo **odds ratio (OR)**, puedo convertirlo en probabilidad usando la fórmula:

$$\text{OR} = \frac{e^{\beta_j}}{e^{\beta_0}}$$

$$\text{OR} = \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}}$$

$$\text{OR} = \frac{\frac{p_1}{1-p_1}}{e^{\beta_0}}$$

$$\text{OR} * e^{\beta_0} = \frac{p_1}{1 - p_1}$$

$$p = \frac{\text{OR} * e^{\beta_0}}{[1 + (\text{OR} * e^{\beta_0})]}$$



# Problema de estudio

Trabajaremos con la información de este artículo. [link](#)

*Ecology of*

FRESHWATER FISH



ORIGINAL ARTICLE

Red operculum spots, body size, maturation and evidence for a satellite male phenotype in non-native European populations of pumpkinseed *Lepomis gibbosus*

Grzegorz Zięba, Carl Smith, Michael G. Fox, Stan Yavno, Eva Záhorská, Miroslaw Przybylski, Gérard Masson, Julien Cucherousset, Hugo Verreycken, Hein H. van Kleef, Gordon H. Copp✉

First published: 23 March 2018 | <https://doi.org/10.1111/eff.12399> | Citations: 1

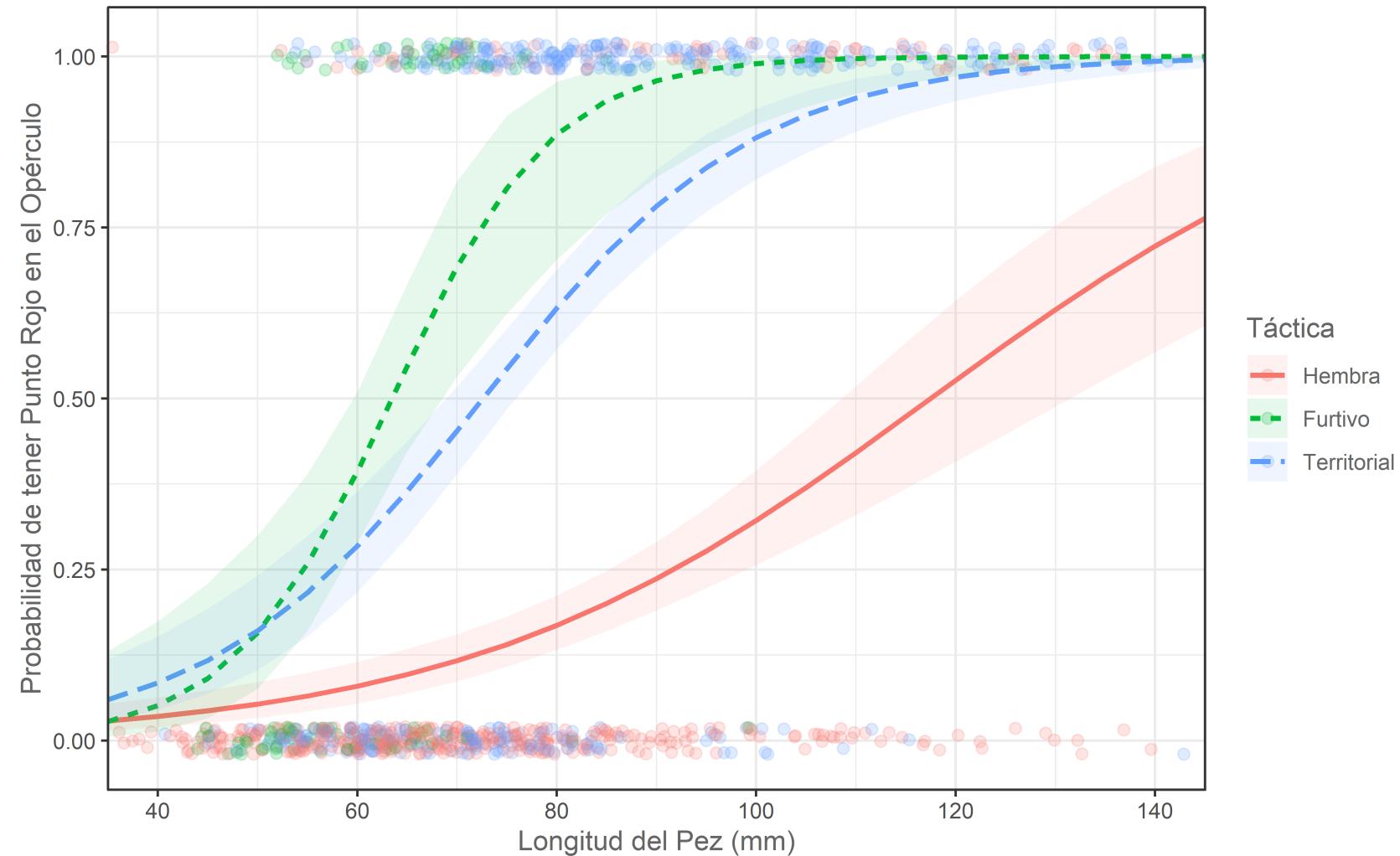


Algunos individuos de los peces *Lepomis gibbosus* tienen un punto rojo sobre su opérculo, el que ha sido asociado con comportamientos de dominancia sobre estrategias de apareamiento. Estas estrategias relacionadas al sexo son:

- Ser **Macho Territorial** y mantener un grupo de nidos.
- Ser **Macho Furtivo** y fecundar nidos de peces territoriales.
- Algunas **hembras** tienen punto rojo, se incluyen en la comparación.



## Gráfico



# Desarrollemos la práctica

60 : 00



# Cuando Y no es Binaria

- Existen casos en los que la variable respuesta no es binaria (no se obtiene de un evento de Bernoulli): y obtenemos **Número de éxitos de un total de N intentos o eventos.**
- Estos son casos donde la distribución de probabilidades es Binomial y no Bernoulli puro.

1. La fertilización cruzada de un Zorro de secura *Lycalopex sechurae* de Ecuador con uno de Perú produce algunas veces zorros totalmente grises. Se quiere encontrar la probabilidad de que nazcan zorros grises de un total de las cinco crías.

$$Y = \# \text{ de zorros grises en cinco crías}$$



# Problema de estudio

Deseamos conocer si los genes 208S2 y AD922 tienen relación con la probabilidad de encontrar un zorro gris en 10 crías.

```
# En estos casos, el modelo sería ( $T = \text{alumbramientos}$ ,  $N = \text{crías}$ )
glm(`(N/T)` ~ x1 + x2, data=DF, family = binomial(link = "logit"), `weights = T`)
```

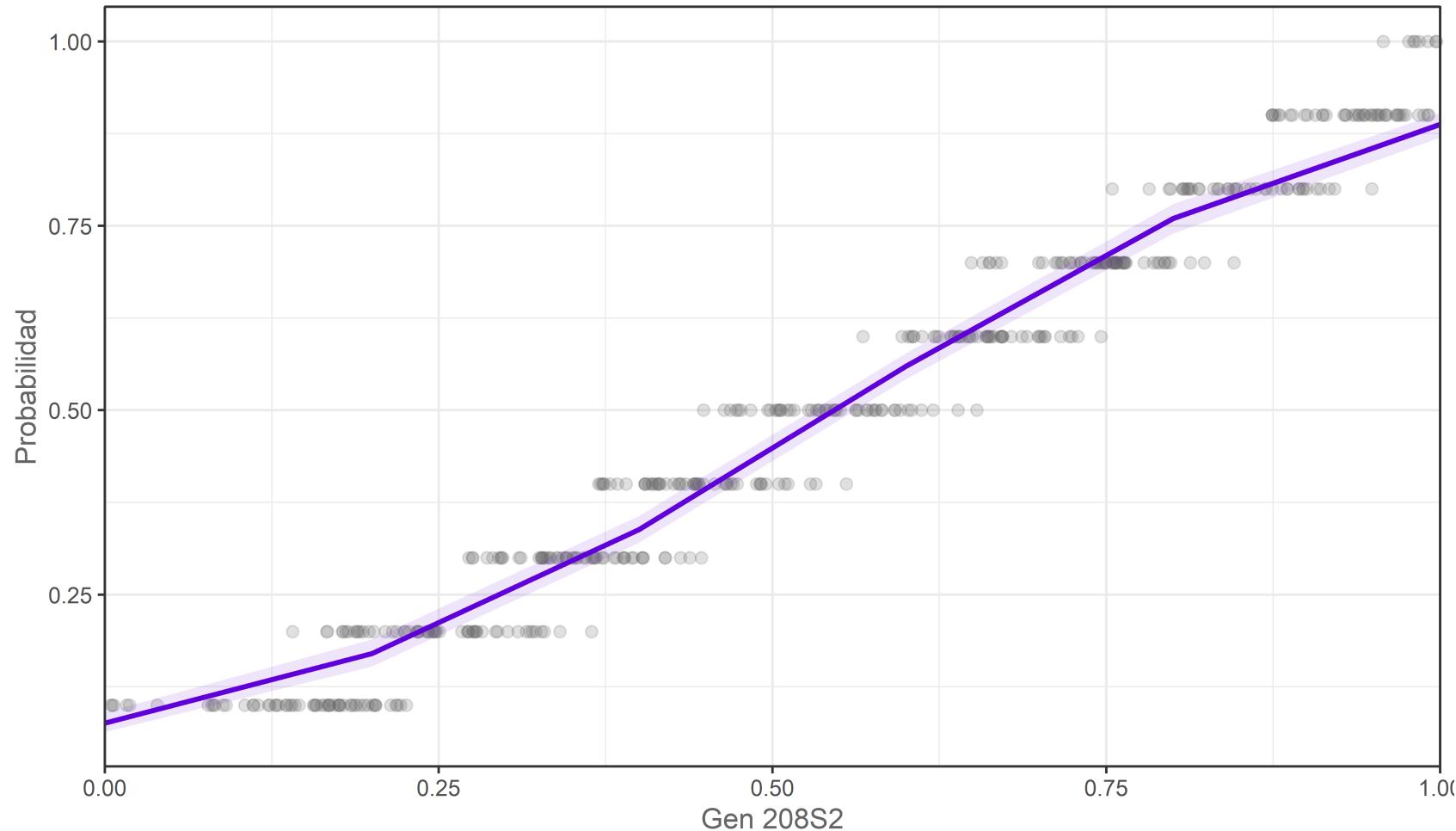
- Variable respuesta:  $Y$  # de crías grises por cada 10 alumbramientos.
- Variables respuesta: genes 208S y AD92.



# Gráfico

## Predicción de la probabilidad de obtención de crías grises

por la expresión del gen 208S2 en poblaciones de Ecuador y Perú de Zorro de Sechura *Lycalopex sechurae*



# Desarrollemos la práctica

30 : 00



# Regresión de Poisson

[ Modelamiento de respuestas discretas tipo conteo ]

# Regresión de Poisson

- Es adecuado para datos ecológicos en los que la variable de respuesta comprende datos de **recuento**:
  - número de individuos.
  - número de especies en un hábitat específico, etc.
- Los datos son **números enteros positivos**.
- Los ceros tienen mucha influencia. **Debe haber pocos ceros**.
- Se asume que la **varianza es aproximadamente igual a la media** (parámetro  $\lambda$ ).



# Caso de estudio Regresión de Poisson:

**Ecology and Evolution**

Open Access

ORIGINAL RESEARCH |  Open Access |  

**Seasonality in spatial distribution: Climate and land use have contrasting effects on the species richness of breeding and wintering birds**

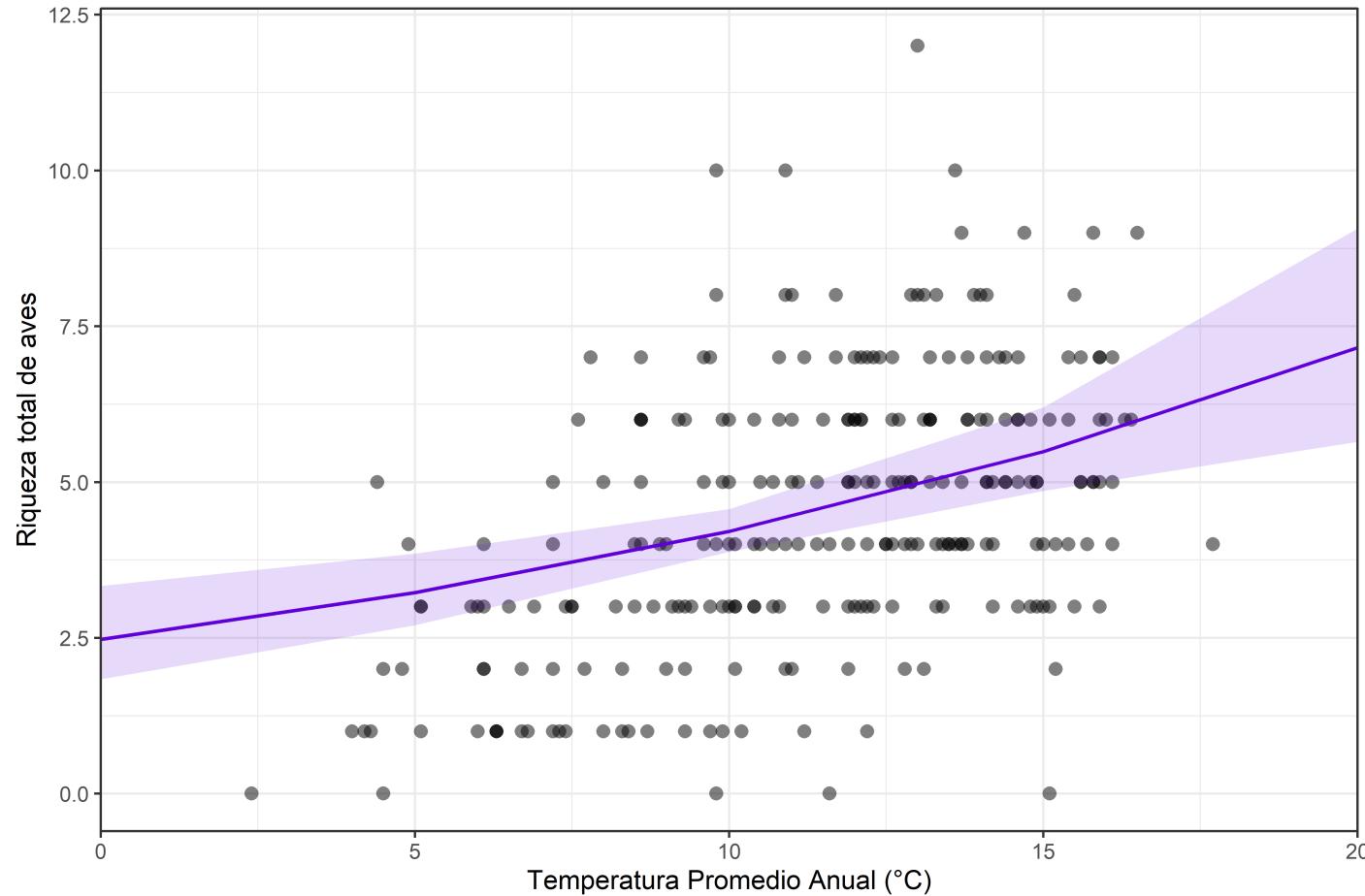
Kazuhiro Kawamura , Yuichi Yamaura, Masayuki Senzaki, Mutsuyuki Ueta, Futoshi Nakamura

First published: 20 June 2019 | <https://doi.org/10.1002/ece3.5286> | Citations: 4

- Base de Datos - Repo Figshare.com
- Artículo Original - Journal Ecology and Evolution
- Datos de **riqueza de especies de aves** (todas las especies, residentes, migrantes de corta distancia y larga distancia tanto en bosques como en pastizales) y **datos ambientales** (temperatura media anual, profundidad de la nieve, elevación y extensión del hábitat circundante) en cada sitio en cada temporada



Gráfico 1



Código 1

```
plot_model(fit, type = "pred", show.data = T  
          terms = c("temp_pa"), show.legend = F)
```

# Desarrollemos la práctica

60 : 00



# Intepretar los coeficientes (R. Poisson)

Para interpretar los resultados debemos exponenciar los coeficientes del modelo:

- $\exp(\beta_0)$  : coeficiente del intercepto. Es el  $\mathbb{E}[Y]$  (es decir el promedio  $\mu$  esperado) cuando todas las  $X_j = 0$ .
- $\exp(\beta_j)$  : coeficiente de la variable  $X_j$ . Con cada unidad de incremento de la variable  $X_j$ , dicha variable tiene un efecto multiplicativo de  $\exp(\beta_j)$  sobre el  $\mathbb{E}[Y]$ .

Interpretar los resultados a primera vista nos puede dar una idea del efecto de  $X_j$  sobre  $\mathbb{E}[Y]$

- Si:  $\beta_j = 0 \rightarrow e^{\beta_j} = 1 \rightarrow$  efecto nulo. El  $\mathbb{E}[Y]$  es  $\beta_0$ . Además,  $Y$  y  $X_j$  no están relacionados.
- Si:  $\beta_j > 0 \rightarrow e^{\beta_j} > 1 \rightarrow$  incremento del conteo en  $\exp(\beta_j)$  veces.
- Si:  $\beta_j < 0 \rightarrow e^{\beta_j} < 1 \rightarrow$  disminución del conteo en  $\exp(\beta_j)$  veces.



# Regresión Binomial Negativa

[ GLM Poisson cuando se detecta sobredispersión ]



# Lidiando con sobredispersión en modelos de conteo

- Los modelos de poisson asumen que  $\frac{\mu}{var} = 0$
- Cuando se detecta sobredispersión, se debe trabajar con modelos Binomiales Negativos.

```
glm_bn <- MASS::glm.nb(y ~ x1 + x2, data=DF, link = "log")
summary(glm_bn)
```



# Gracias por su atención

