

# **Clase 18 Introducción a Modelos Lineales Generales**

**Diplomado en Análisis de Datos con R e Investigación  
reproducible para Biociencias.**

Dr. José Gallardo Matus y Dra. María Angélica Rueda

Pontificia Universidad Católica de Valparaíso

08 November 2022

# PLAN DE LA CLASE

## 1.- Introducción

- ▶ Modelos lineales generales ¿Qué son y para que sirven?
- ▶ Ejemplos de modelos lineales generales.
- ▶ Interpretación de MLG con R.

## 2.- Práctica con R y Rstudio cloud

- ▶ Ajustar modelos lineales generales.
- ▶ Realizar gráficas avanzadas con ggplot2.
- ▶ Elaborar un reporte dinámico en formato html.

# INTRODUCCIÓN

Durante años, los modelos lineales clásicos (normales) han sido usados como la metodología de análisis a la hora de intentar describir la mayoría de los fenómenos que ocurren en el entorno.

**¿Qué podemos hacer cuando los datos no se ajustan a un modelo lineal?**

- ▶ Muchas veces se recurre a transformar la variable respuesta.
- ▶ Pero al aplicar la transformación a la variable respuesta, **NO** necesariamente se cumplirían todos los supuestos.
- ▶ Las interpretaciones deben hacerse en términos de la **variable transformada**.

# ¿QUÉ SON LOS MODELOS LINEALES GENERALES?

Los modelos lineales generales extienden a los modelos lineales clásicos admitiendo distribuciones no normales para la variable respuesta y modelando funciones de la media.

Los MLG incluyen como casos particulares a los siguientes modelos:

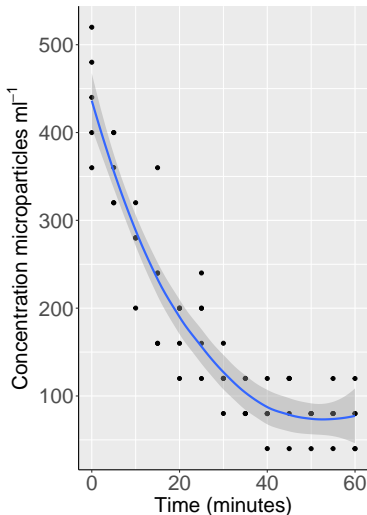
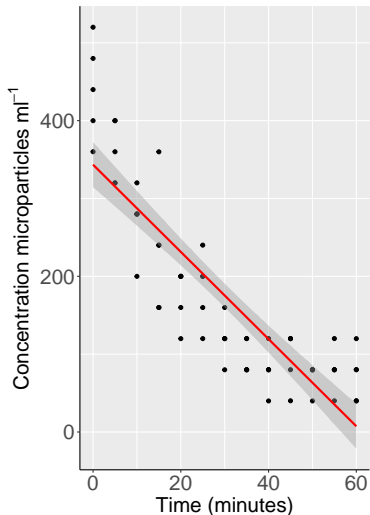
- ▶ Modelos Lineales Clásicos: **Modelo de regresión lineal simple, modelo de regresión lineal múltiple, ANOVA , ANCOVA.**
- ▶ Modelos no lineales (con variables predictoras elevadas a alguna potencia (cuadráticas, cúbicas, etc)).
- ▶ Modelo de regresión logística.

# ¿POR QUÉ USAR MODELOS LINEALES GENERALES?

- ▶ Modelos que reflejan mejor la naturaleza de los datos.
- ▶ Hay variables respuestas que son **resistentes** a ser transformadas (**por ej.** Variables discretas, o variables con gran cantidad de ceros).
- ▶ Las relaciones lineales generalmente fuerzan las predicciones del espacio de la variable respuesta (**por ej.** Predicción de valores negativos cuando la variable respuesta es un conteo).

# REGRESIÓN NO LINEAL CUADRÁTICA

En este ejemplo vamos a comparar el modelo lineal vs. el modelo no lineal con término cuadrático.



# MODELO LINEAL

## Modelo 1:

$$\text{log\_microparticle\_concentration} = \beta_0 + \beta_1 \text{time} + \epsilon$$

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.57	0.03	76.97	0
time	-0.01	0.00	-14.96	0

$$R^2 = 0.78, p\text{-val} = 2.0490325 \times 10^{-22}$$

# MODELO NO LINEAL (INCLUYE TÉRMINO CUADRÁTICO)

Modelo 2:

$$\text{log\_microparticle\_concentration} = \beta_0 + \beta_1 \text{time} + \beta_2 \text{time}^2 + \epsilon$$

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.14	0.02	130.92	0
poly(time, 2)1	-2.13	0.13	-16.13	0
poly(time, 2)2	0.44	0.13	3.35	0

$$R^2 = 0.81, p\text{-val} = 2.2610223 \times 10^{-23}$$



# COMPARACIÓN DE MODELOS

► Modelo 1:

$$\text{log\_microparticle\_concentration} = \beta_0 + \beta_1 \text{time} + \epsilon$$

► Modelo 2:

$$\text{log\_microparticle\_concentration} = \beta_0 + \beta_1 \text{time} + \beta_2 \text{time}^2 + \epsilon$$

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
63	1.275337	NA	NA	NA	NA
62	1.080344	1	0.194993	11.19047	0.0013997

# REGRESIÓN LOGÍSTICA

- ▶ La regresión logística es útil para predecir variables respuesta de naturaleza binaria: Presencia o ausencia, sano o enfermo, maduro o no maduro, macho o hembra.
- ▶ Las principales supuestos del modelo de regresión logística son:
  - a) Respuesta binaria: La variable respuesta debe ser binaria.
  - b) Independencia: las observaciones deben ser independientes.
  - c) Multicolinealidad: se requiere de muy poca a ninguna multicolinealidad entre los predictores (para regresión logística múltiple).
  - d) Linealidad: entre la variable independiente y el logaritmo natural de la variable respuesta.

# REGRESIÓN LOGÍSTICA SIMPLE

Modelo de regresión logística simple:

$$p(Y = 1) = \frac{e^{\beta_0 + \beta_1 X_1}}{1 + e^{\beta_0 + \beta_1 X_1}}$$

$p(Y = 1)$  = Probabilidad de que la variable respuesta dicotómica tome un valor de 1 (éxito).

$X_1$  = Variable predictora.

$B_0$  = Intercepto.

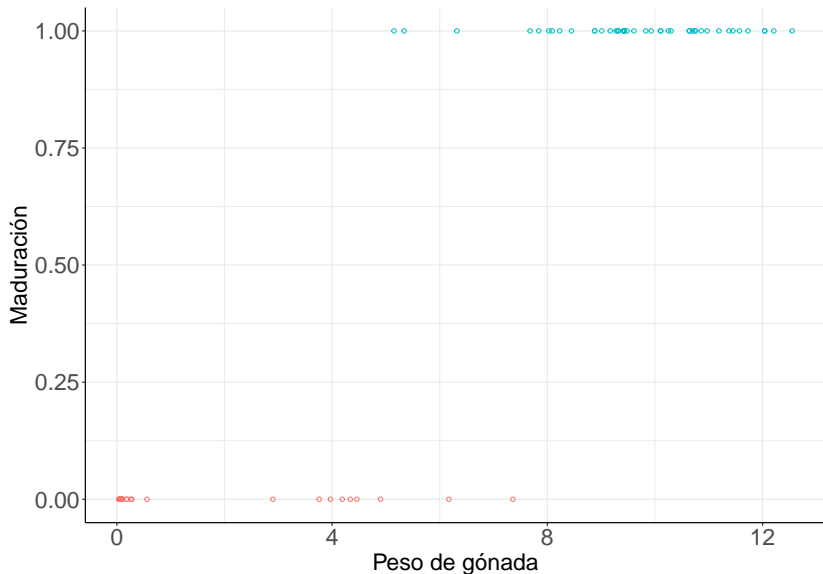
$B_1$  = Pendiente.

# ESTUDIO DE CASO 2: MADURACIÓN EN SALMÓN DEL ATLÁNTICO

En este estudio de caso trabajaremos con un subconjunto de la base de datos relacionada a la maduración en salmones machos ( $n=90$ ).

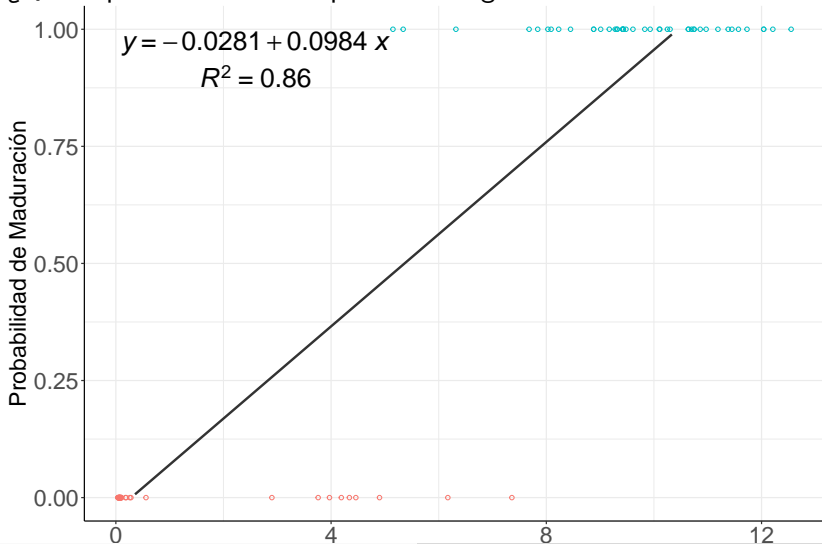
variable	Descripción
<b>Fish</b>	Identificador del salmón
<b>Genotype</b>	Genotipo
<b>Gonad</b>	Peso de gónada
<b>Maturation</b>	estado de maduración (1: maduro) o (0: inmaduro)

# RELACIÓN ENTRE MADURACIÓN VS PESO DE GÓNADA



# REGRESIÓN LINEAL ENTRE MADURACIÓN VS PESO DE GÓNADA

¿Qué supuestos no se cumplen de la regresión lineal?



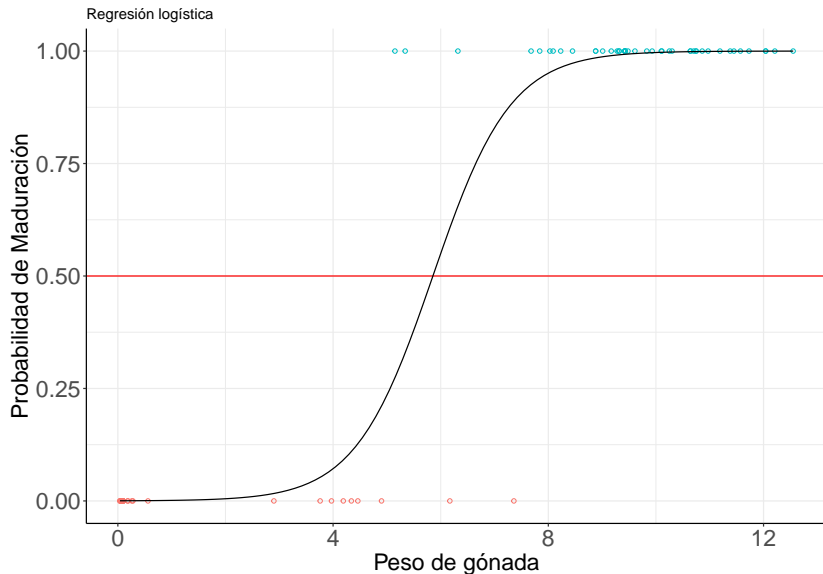
# MODELO LINEAL

$$\text{Maduración} = \beta_0 + \beta_1 \text{ Peso de gónada} + \epsilon$$

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.03	0.03	-0.92	0.36
Gonad	0.10	0.00	22.89	0.00

$$R^2 = 0.86, p\text{-val} = 7.977942 \times 10^{-39}$$

# PREDICCIÓN LOGÍSTICA





# REGRESIÓN LOGÍSTICA SIMPLE

```
mod_logit <- glm(Maturation ~ Gonad,  
                 family= binomial, data = maduracion)  
summary(mod_logit)$coef %>% kable(digits = 2)
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-8.09	2.64	-3.06	0
Gonad	1.38	0.43	3.25	0

# EJEMPLO PREDICCIÓN DE AMBOS MODELOS

Probabilidad de estar maduro para un peso de gónada de 4 gramos.

## REGRESIÓN LINEAL

Probabilidad de maduración
0.3656176

## [1] "No madura"

## REGRESIÓN LOGÍSTICA

Probabilidad de maduración
0.0715492

## [1] "No madura"

# RELACIÓN ENTRE MADURACIÓN VS GENOTIPO

Genotipo E = Maduración temprana o Early.

Genotipo L = Maduración tardía o Late.

¿Qué genotipo tiene mayor probabilidad de maduración?

```
table(maduracion$Maturation, maduracion$Genotype) %>%  
  kable()
```

	EE	EL	LL
0	4	22	19
1	44	1	0

# REGRESIÓN LOGÍSTICA MÚLTIPLE

Modelo de regresión logística múltiple:

$$p(Y = 1) = \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}$$

$p(Y = 1)$  = Probabilidad de que la variable respuesta dicotómica tome un valor de 1 (éxito).

$X_1$  = Variable predictora 1.

$X_p$  = Variable predictora  $p$ .

$B_0$  = Intercepto.

$B_1$  = Pendiente variable  $X_1$ .

$B_p$  = Pendiente variable  $X_p$ .

# REGRESIÓN LOGÍSTICA MÚLTIPLE

```
mod_logit_mult <- glm(Maturation ~ Gonad +  
                      Genotype,family= binomial,  
                      data = maduracion)  
summary(mod_logit_mult)$coef %>% kable(digits = 2)
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-5.95	3.16	-1.88	0.06
Gonad	1.14	0.45	2.50	0.01
GenotypeEL	-1.30	1.65	-0.78	0.43
GenotypeLL	-16.85	3447.62	0.00	1.00

# REGRESIÓN LOGÍSTICA (MODELO NULO)

```
mod_nulo <- glm(Maturation ~ 1,  
                family= binomial, data = maduracion)  
summary(mod_nulo)$coef %>% kable(digits = 2)
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0	0.21	0	1

# COMPARACIÓN DE MODELOS AIC Y BIC

```
AIC(mod_nulo,mod_logit,mod_logit_mult)%>% kable()
```

	df	AIC
mod_nulo	1	126.76649
mod_logit	2	18.30228
mod_logit_mult	4	21.25087

```
BIC(mod_nulo,mod_logit,mod_logit_mult)%>% kable()
```

	df	BIC
mod_nulo	1	129.26630
mod_logit	2	23.30190
mod_logit_mult	4	31.25011

# COMPARACIÓN DE MODELOS (ANOVA)

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
89	124.76649	NA	NA	NA
88	14.30228	1	110.464210	0.0000000
86	13.25087	2	1.051411	0.5911383



# RESUMEN DE LA CLASE

- 1). Revisión de conceptos: modelos lineales generales.
- 2). Construir y ajustar modelos lineales generales.