Clase 15 Introducción a Modelos Lineales Generales

OCE 386 - Introducción al análisis de datos con R.

Dr. José A. Gallardo y y Dra. María Angélica Rueda | Pontificia Universidad Católica de Valparaíso

23 November 2021

PLAN DE LA CLASE

1.- Introducción

- Modelos lineales generales ¿Qué son y para que sirven?
- Regresión cuadrática.
- Regresión logística.
- 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 pdf.

INTRODUCCIÓN

Los modelos **lineales** clásicos permiten describir la mayoría de los fenómenos que ocurren en el entorno, siempre que la relación entre variables sea lineal.

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

- Muchas veces se recurre a transformar la variable respuesta (Logaritmo).
- Pero la transformación de la variable respuesta **NO** necesariamente permite cumplir con todos los supuestos.
- Las interpretaciones deben hacerse en términos de la **variable transformada**.

¿QUÉ SON LOS MODELOS LINEALES GENERALES (MLG)?

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

Los MLG incluyen como casos particulares a los siguientes modelos:

- Modelos Lineales: **Regresión lineal simple, regresión lineal múltiple**
- Modelos no lineales: Con variables predictoras elevadas a alguna potencia (cuadráticas, cúbicas, etc).
- Modelo de regresión logística: Variable respuesta binaria.

¿POR QUÉ USAR MODELOS LINEALES GENERALES?

- 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).

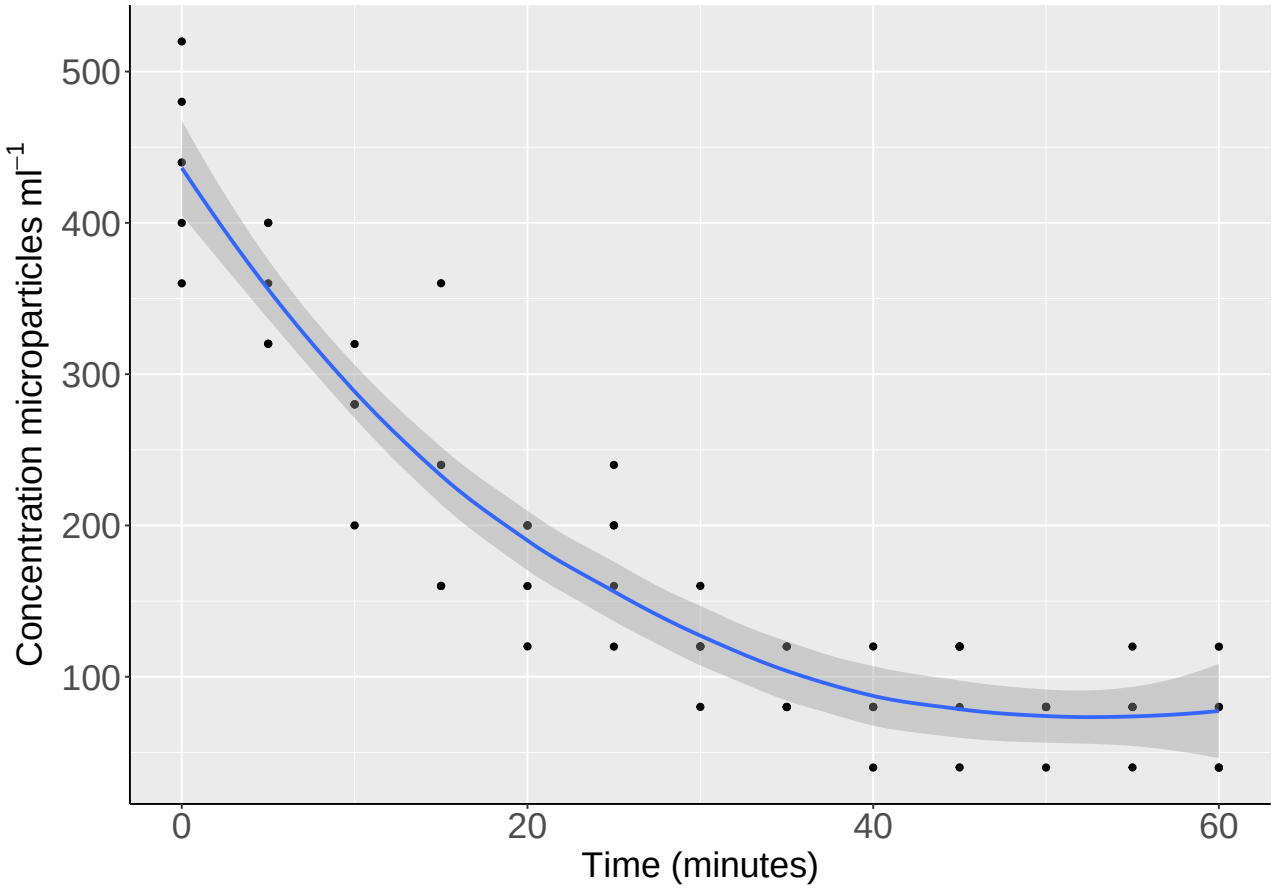
ESTUDIO DE CASO TASA DE ACLARACIÓN EN MITILIDOS

Tasa de aclaración dieta artificial en mitilidos.
Fuente: Willer and Aldridge 2017

time	sample	replicate	particle concentration
0	mussel	a	400
5	mussel	a	320
10	mussel	a	280
...	...	...	...
0	control	a	160
5	Control	a	120
10	Control	a	120

TASA DE ACLARACIÓN MUSSEL.

Problemas: La concentración es discreta y la relación no es lineal.



Tips: stat_smooth(method='loess',formula=y~x, se=T)

MODELO LINEAL

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

Modelo 1:

Log (Microparticle concentration) $= \beta_0 + \beta_1 time + \epsilon$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.567087	0.0333508	76.97221	0
time	-0.014116	0.0009433	-14.96447	0

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

MODELO NO LINEAL (INCLUYE TÉRMINO CUADRÁTICO)

Modelo 2:

Microparticle concentration $= \beta_0 + \beta_1 time + \beta_2 time^2 + \epsilon$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.1436057	0.0163730	130.923107	0.0000000
poly(time, 2)1	-2.1291367	0.1320034	-16.129403	0.0000000
poly(time, 2)2	0.4415801	0.1320034	3.345217	0.0013997

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

COMPARACIÓN DE MODELOS

- Modelo 1:

Log(microparticle concentration) $= \beta_0 + \beta_1 time + \epsilon$

- Modelo 2:

Microparticle concentration $= \beta_0 + \beta_1 time + \beta_2 time^2 + \epsilon$

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

Las principales condiciones de la regresión logística son:

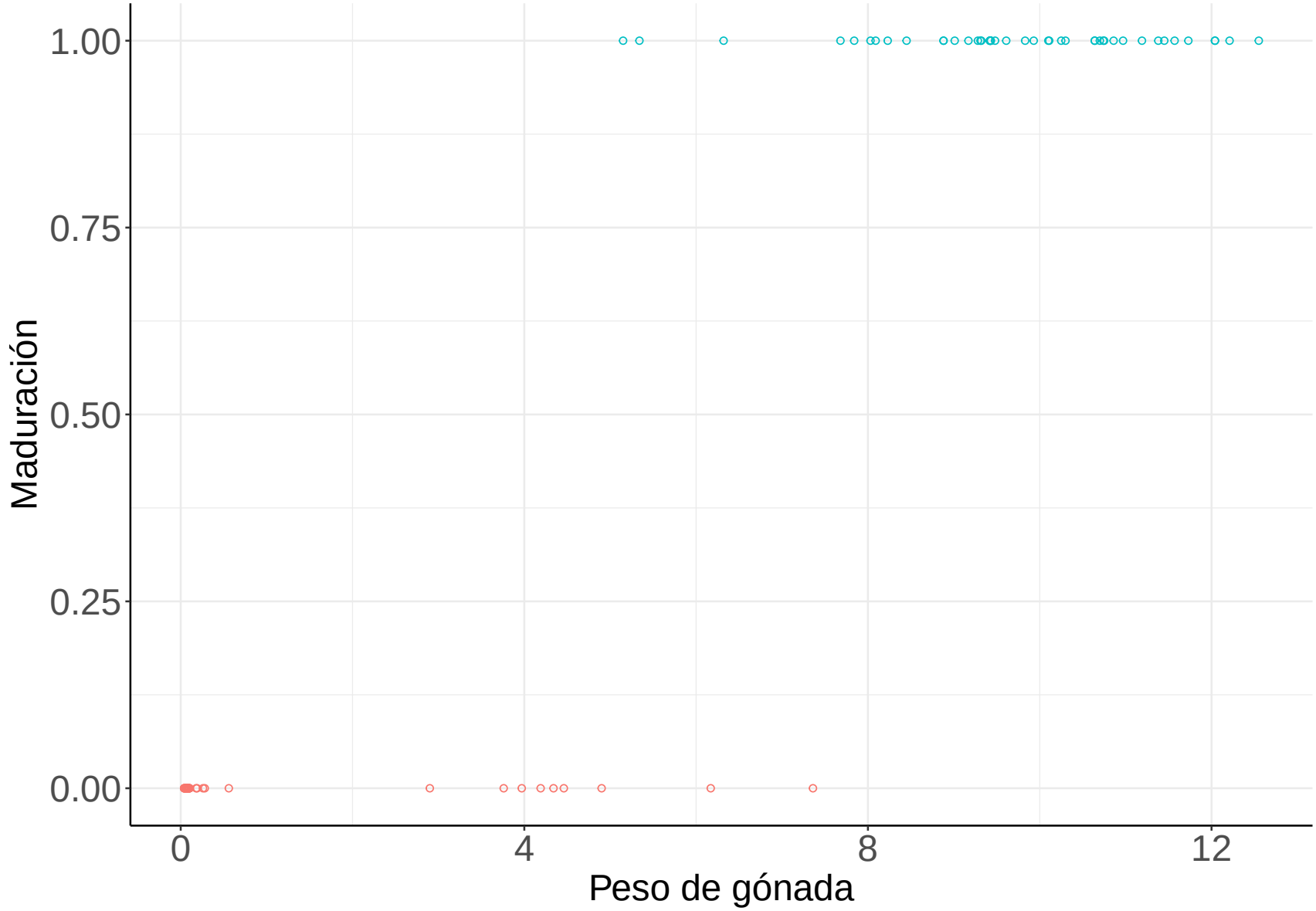
- Respuesta binaria: La variable respuesta debe ser binaria.
- Independencia: las observaciones deben ser independientes.
- Multicolinealidad: se requiere de muy poca a ninguna multicolinealidad entre los predictores (para regresión logística múltiple).
- Linealidad: entre la variable independiente y el logaritmo natural de odds (Cociente de chances).

ESTUDIO DE CASO: MADURACIÓN EN SALMÓN DEL ATLÁNTICO

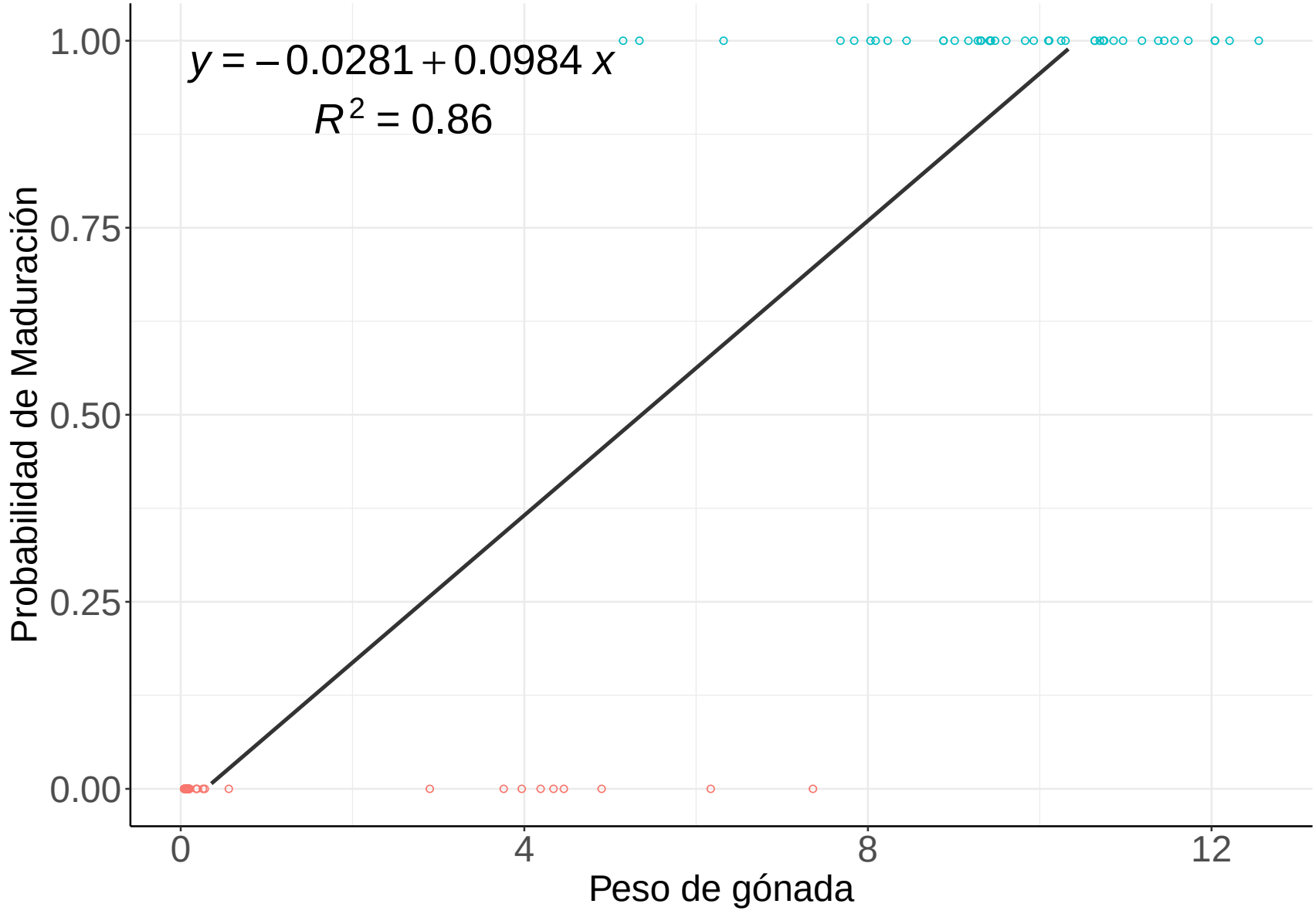
Estudio de la relación entre peso de la gónada y nivel de maduración en salmones (n=90, solo machos).

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

RELACIÓN ENTRE MADURACIÓN VS PESO DE GÓNADA



RELACIÓN LINEAL ENTRE MADURACIÓN VS PESO DE GÓNADA



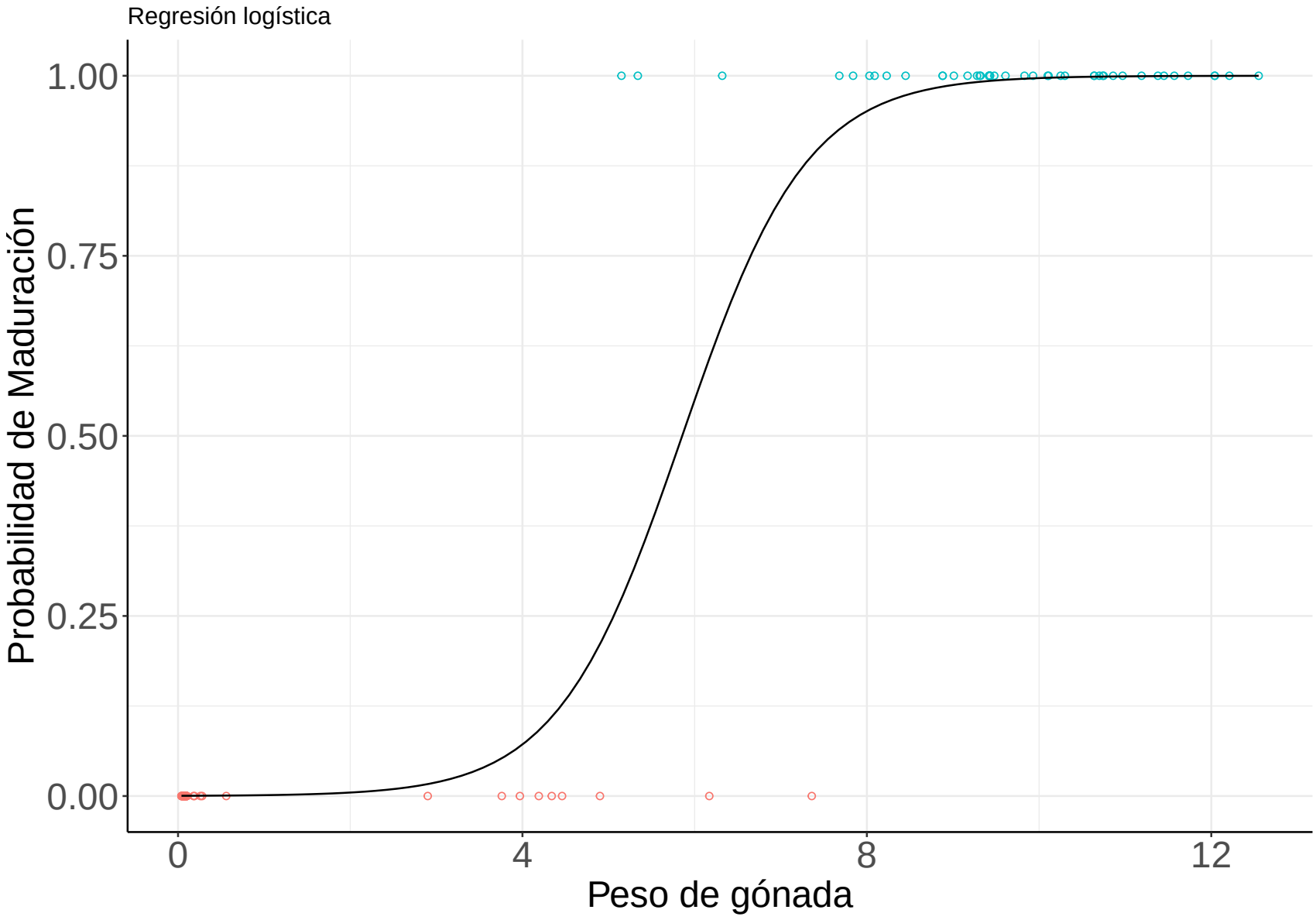
MODELO LINEAL

Maduración $= \beta_0 + \beta_1$ Peso de gónada $+ \epsilon$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0280808	0.0306710	-0.9155493	0.3624054
Gonad	0.0984246	0.0042997	22.8908036	0.0000000

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

RELACIÓN SIGMOIDEA ENTRE MADURACIÓN VS PESO DE GÓNADA



PREDECIR SI UN SALMÓN MADURA O NO PARA UN PESO DE GÓNADA DE 4

CONSIDERANDO LA REGRESIÓN LINEAL

Probabilidad de maduración
0.3656176

```
## [1] "No madura"
```

CONSIDERANDO LA REGRESIÓN LOGÍSTICA

Probabilidad de maduración
0.0715492

```
## [1] "No madura"
```

REGRESIÓN LOGÍSTICA (MODELO NULO)

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

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0	0.2108185	0	1

$$P(X) = \frac{e(\beta_0)}{1 + e(\beta_0)}$$

REGRESIÓN LOGÍSTICA SIMPLE

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

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.089844	2.6425566	-3.06137	0.0022033
Gonad	1.381678	0.4255612	3.24672	0.0011674

$$P(X) = \frac{e(\beta_0 + \beta_1 X)}{1 + e(\beta_0 + \beta_1 X)}$$

REGRESIÓN LOGÍSTICA MÚLTIPLE

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

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.951859	3.1608767	-1.8829772	0.0597035
Gonad	1.135307	0.4546516	2.4970928	0.0125216
GenotypeEL	-1.296134	1.6538041	-0.7837292	0.4331990
GenotypeLL	-16.852220	3447.6185502	-0.0048881	0.9960999

$$P(X) = \frac{e(\beta_0 + \beta_1 X + \beta_2 X)}{1 + e(\beta_0 + \beta_1 X + \beta_2 X)}$$

COMPARACIÓN DE MODELOS POR ANOVA

```
anova(mod_nulo,mod_logit,mod_logit_mult,
      test ='Chisq')%>% kable()
```

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

COMPARACIÓN DE MODELOS POR AIC

```
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

RESUMEN DE LA CLASE

1). Revisión de conceptos: modelos lineales generales.

2). Construir y ajustar modelo de regresión cuadrática.

3). Construir y ajustar modelo de regresión logística.

4). Comparar modelos de regresión.