Clase 09 Regresión Logística

Curso Análisis de datos con R para Biociencias

Dra. María Angélica Rueda. maria.rueda.c@pucv.cl | Pontificia Universidad Católica de Valparaíso

27 January 2022

PLAN DE LA CLASE

1.- Introducción

- Regresión polinomial.
- Modelos de Regresión logística.
- Ejemplo de modelo Regresión logística.
- Interpretación de modelos de regresión con R.

2.- Práctica con R y Rstudio cloud

- Ajustar modelos de regresión logística.
- Realizar gráficas avanzadas con ggplot2.
- Elaborar un reporte dinámico en formato pdf.

REGRESIÓN POLINOMIAL

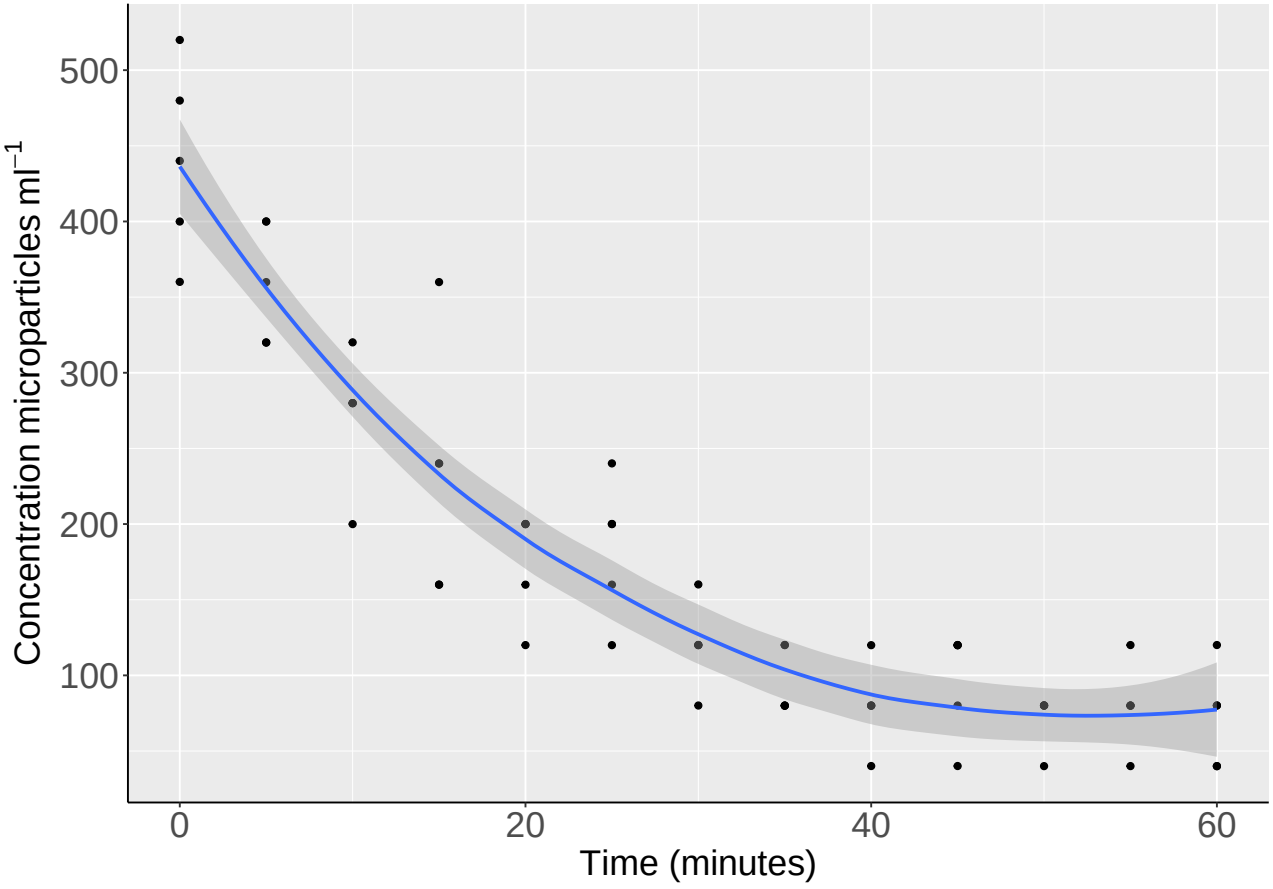
Sea Y una variable respuesta continua y la variable predictora X, un modelo de regresión polinomial se puede representar como,

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta_h X^h + \epsilon$$

donde h es el grado del polinomio.

REGRESIÓN POLINOMIAL

En este ejemplo vamos a comparar la regresión lineal simple con variable linealizada *vs* la regresión polinomial con término cuadrático.



REGRESIÓN LINEAL SIMPLE: RECORDATORIO

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

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

REGRESIÓN POLINOMIAL CON TÉRMINO CUADRÁTICO

Modelo 2:

log_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:

log_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

La regresión logística no requiere de ciertas condiciones como linealidad, normalidad y homocedasticidad de los residuos que sí lo son para la regresión lineal. Las principales condiciones que este modelo requiere son:

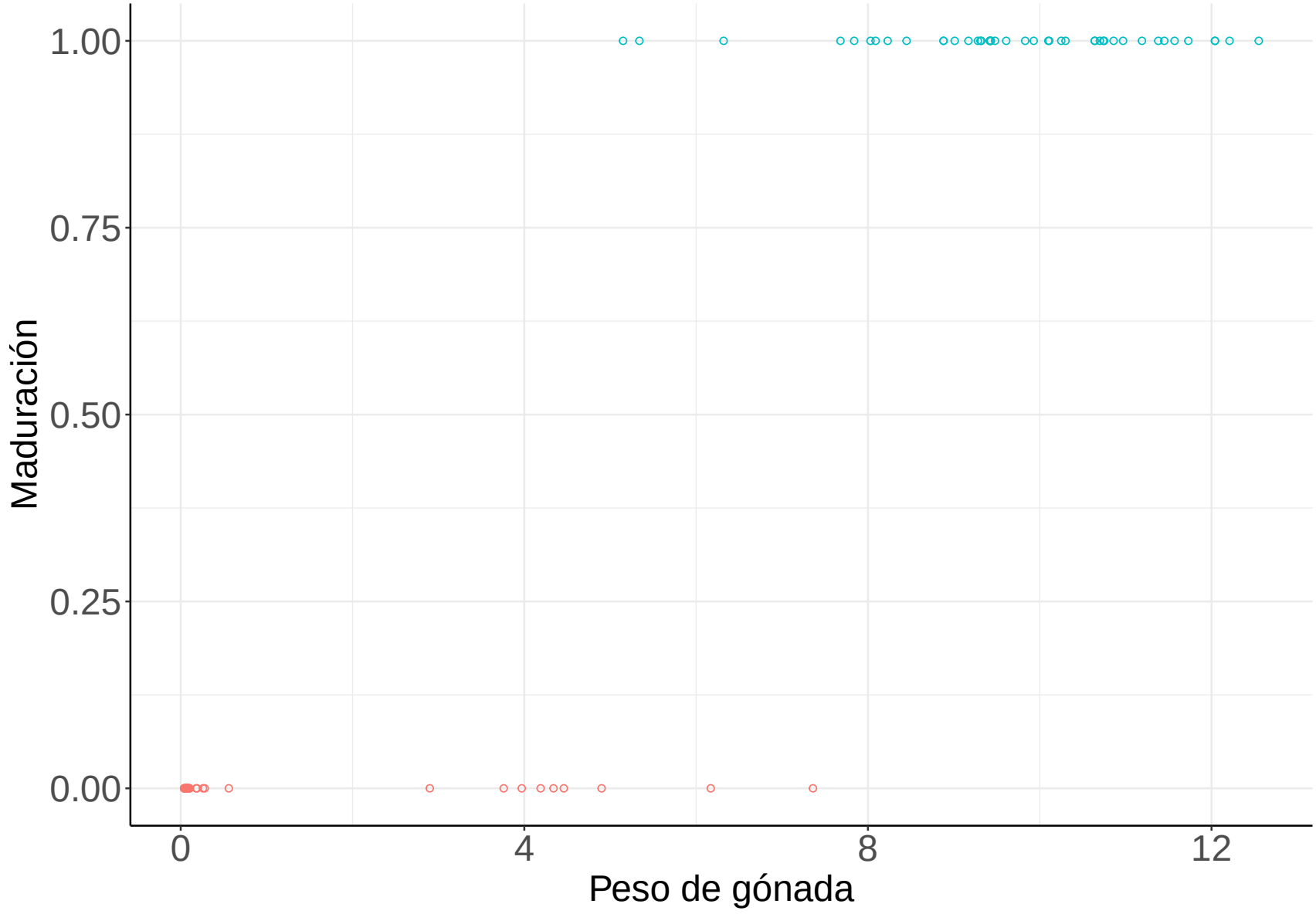
- Respuesta binaria: La variable respuesta debe ser binaria.
- Independencia: las observaciones deben ser independientes.
- Multicolinealidad: se requiere de muy poca a ninguna multicolinealidad entre los predictores (para regresión logística múltiple).
- Linealidad: entre la variable independiente y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

ESTUDIO DE CASO 2: MADURACIÓN EN SALMÓN DEL ATLÁNTICO

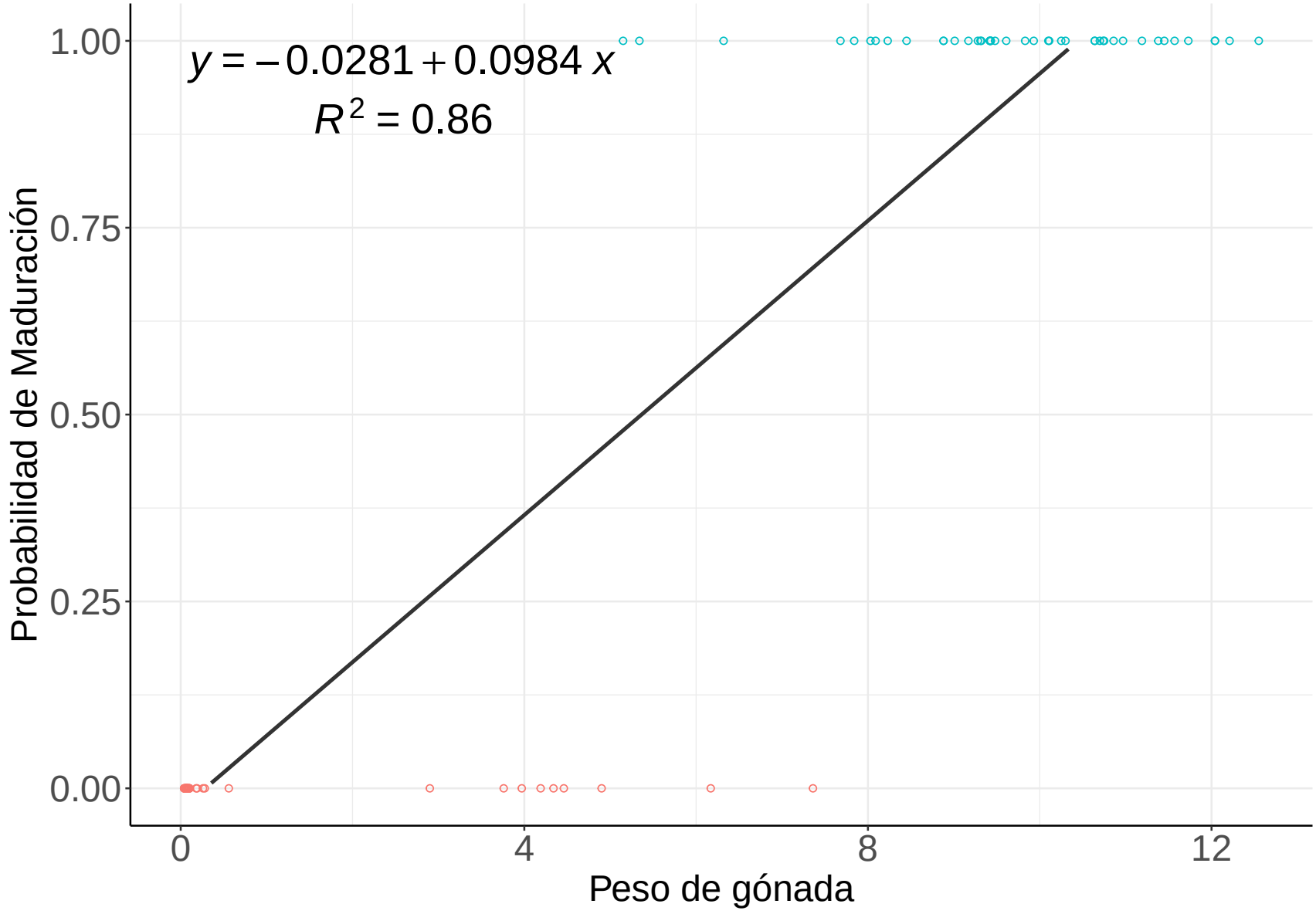
En este estudio de caso trabajaremos con un subconjunto de la base de datos relacionada a la maduración en salmones machos (n=90).

Variable	**Descripción**
Fish	Identificador del salmón
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



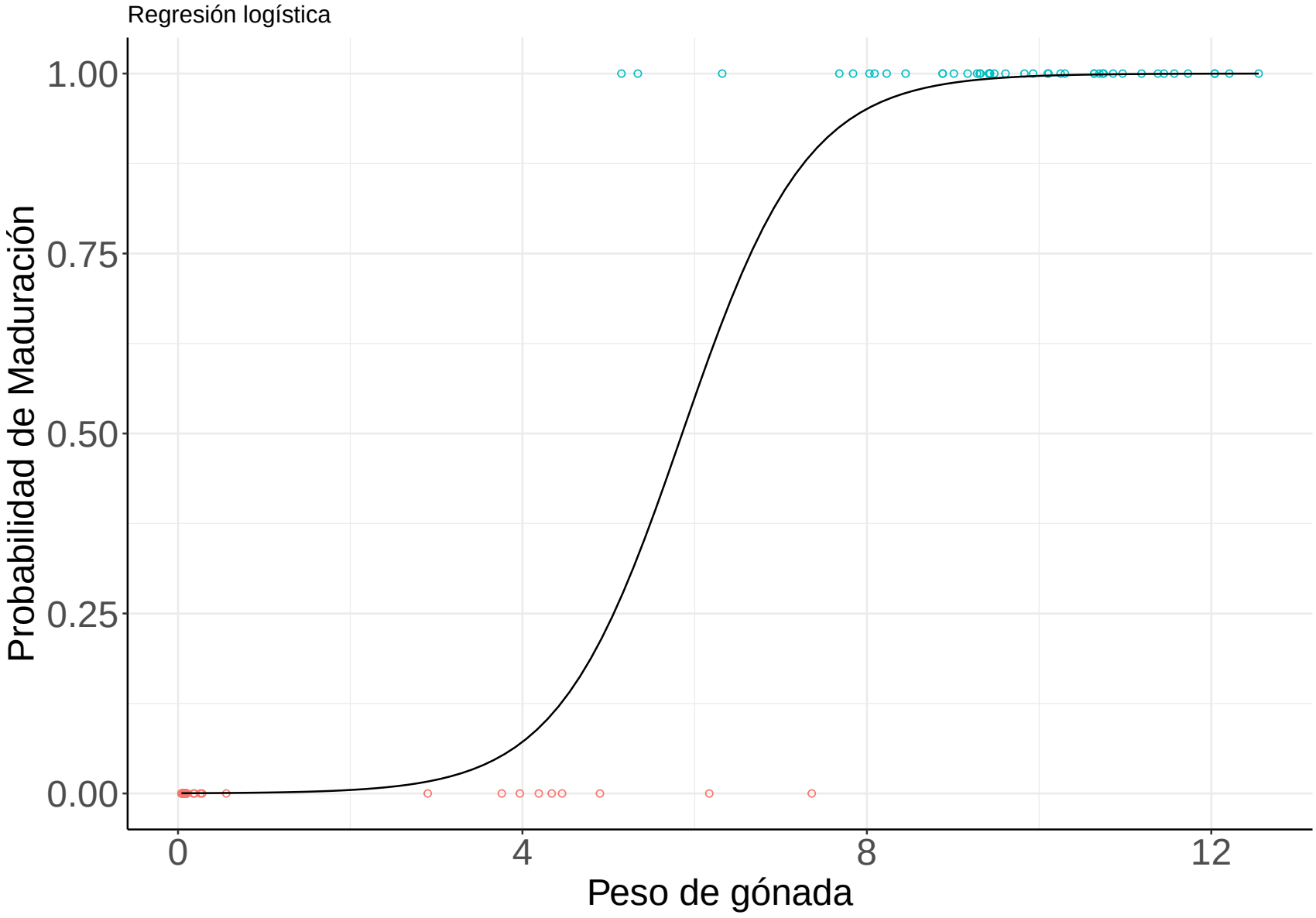
MODELO LINEAL

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

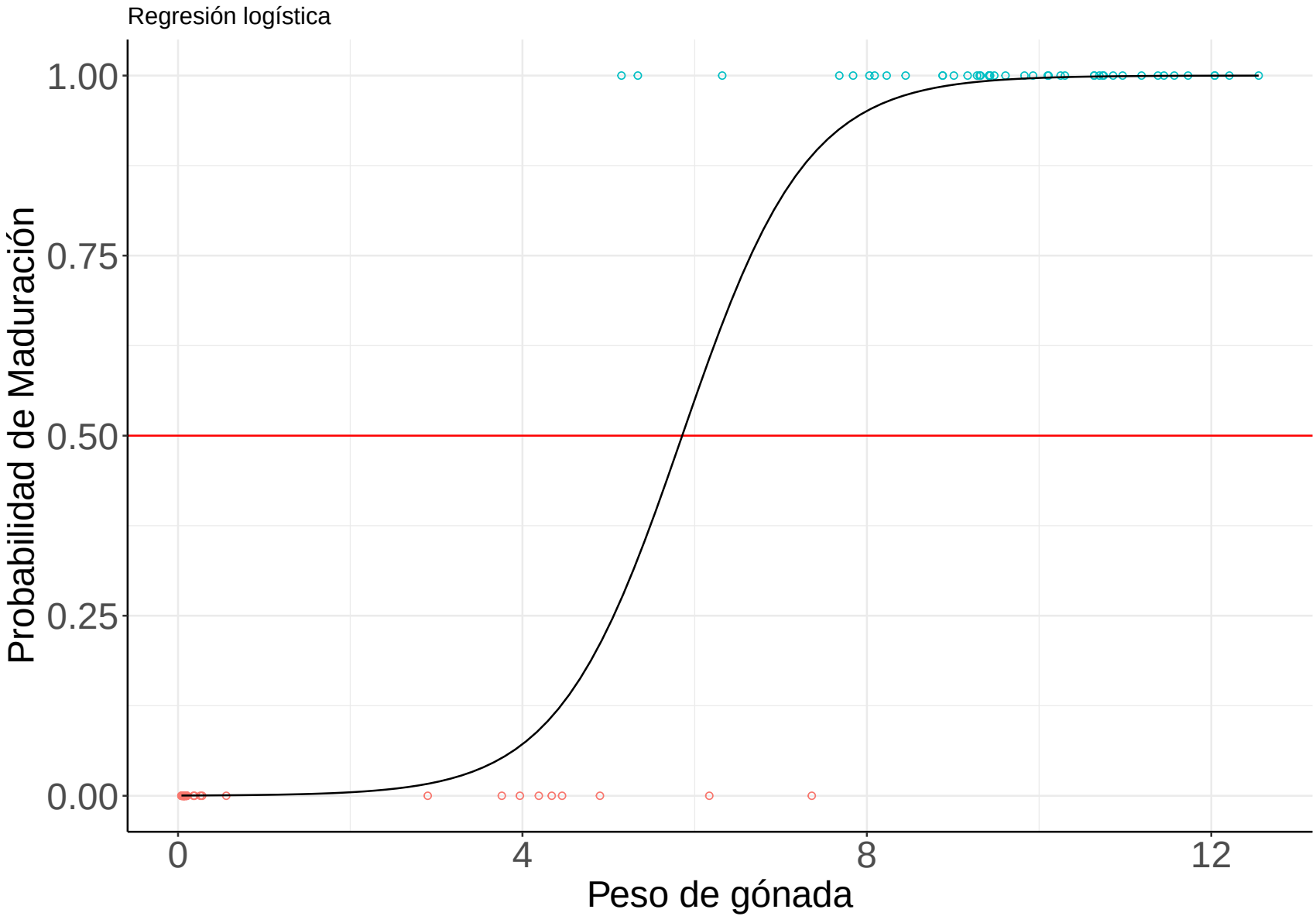
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0280808	0.0306710	-0.9155493	0.3624054
Gonad	0.0984246	0.0042997	22.8908036	0.0000000

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

RELACIÓN SIGMOIDEA ENTRE MADURACIÓN VS PESO DE GÓNADA



PREDICCIÓN MODELO LINEAL VS MODELO NO LINEAL



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

CONSIDERANDO LA REGRESIÓN LINEAL

```
##
##       0  1
##    0 43  2
##    1  2 43
```

Probabilidad de maduración
0.3656176

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

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

CONSIDERANDO LA REGRESIÓN LOGÍSTICA

```
##
##        0  1
##    0 43  2
##    1  2 43
```

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

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

COMPARACIÓN DE MODELOS AIC

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

	df	AIC
mod_nulo	1	126.76649
mod_logit	2	18.30228

COMPARACIÓN DE MODELOS (ANOVA)

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
89	124.76649	NA	NA	NA
88	14.30228	1	110.4642	0

RESUMEN DE LA CLASE

1). Revisión de conceptos: Regresión Logística.

2). Construir y ajustar modelos de Regresión Logística.