Modelos estadísticos II: Modelos lineales generalizados

Técnicas estadísticas avanzadas para la conservación de la biodiversidad - Universidad de Huelva

David García Callejas

01/2021

Modelos lineales generalizados

- Hasta ahora: modelos lineales con variable respuesta **continua** y residuos **normales**

```
gupp <- read.csv(here::here("datasets",
                            "chap17f9_1GuppyColdDeath.csv"))
head(gupp)
```

```
##   fish exposureDurationMin mortality
## 1    1                   3         1
## 2    2                   3         1
## 3    3                   3         1
## 4    4                   3         1
## 5    5                   3         1
## 6    6                   3         1
```

Modelos lineales generalizados

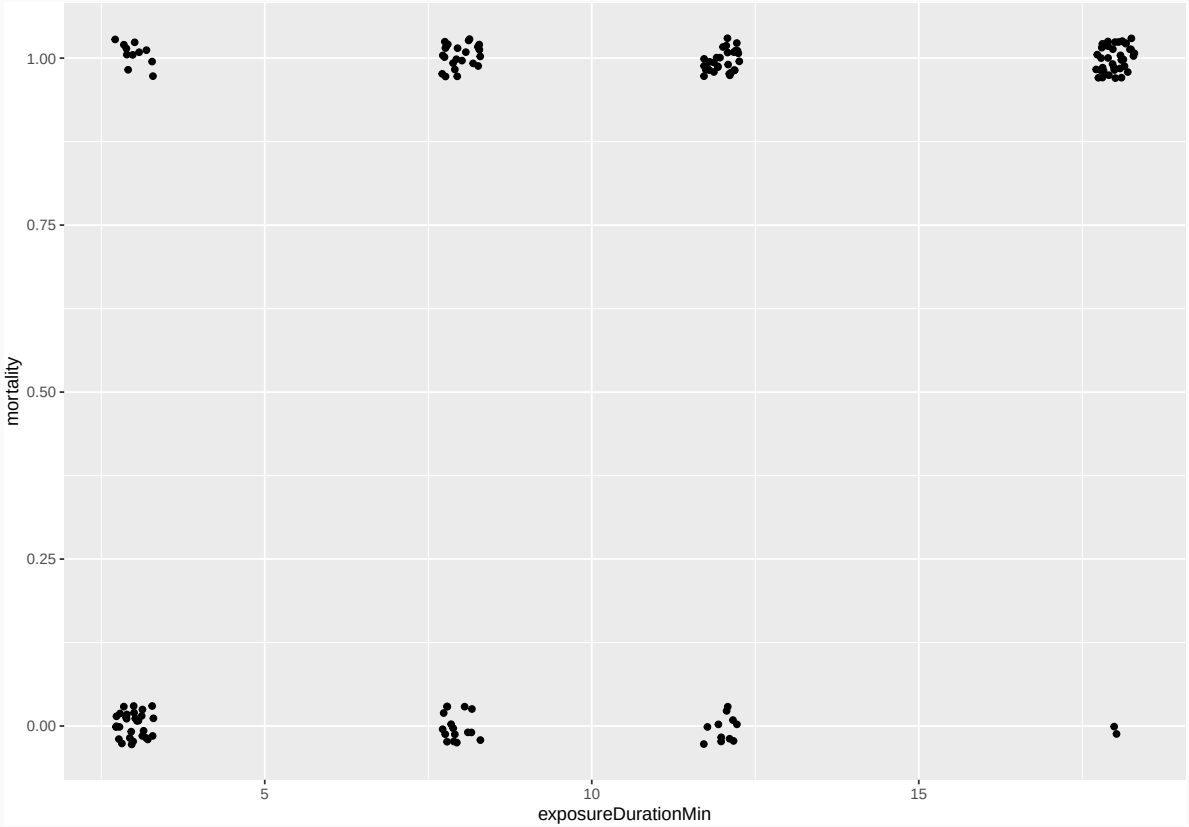
- Hasta ahora: modelos lineales con variable respuesta **continua** y residuos **normales**
- ¿podemos modelar variables con respuestas discretas? Por ejemplo, mortalidad de peces en función de tiempo de exposición a temperaturas de 5ºC:

```
gupp <- read.csv(here::here("datasets",
                            "chap17f9_1GuppyColdDeath.csv"))
head(gupp)
```

```
##   fish exposureDurationMin mortality
## 1    1                   3         1
## 2    2                   3         1
## 3    3                   3         1
## 4    4                   3         1
## 5    5                   3         1
## 6    6                   3         1
```

Modelos lineales generalizados

```
ggplot(gupp, aes(x = exposureDurationMin,y = mortality)) +
  geom_point(position = position_jitter(width = .3,height = .03))
```



Modelos lineales generalizados

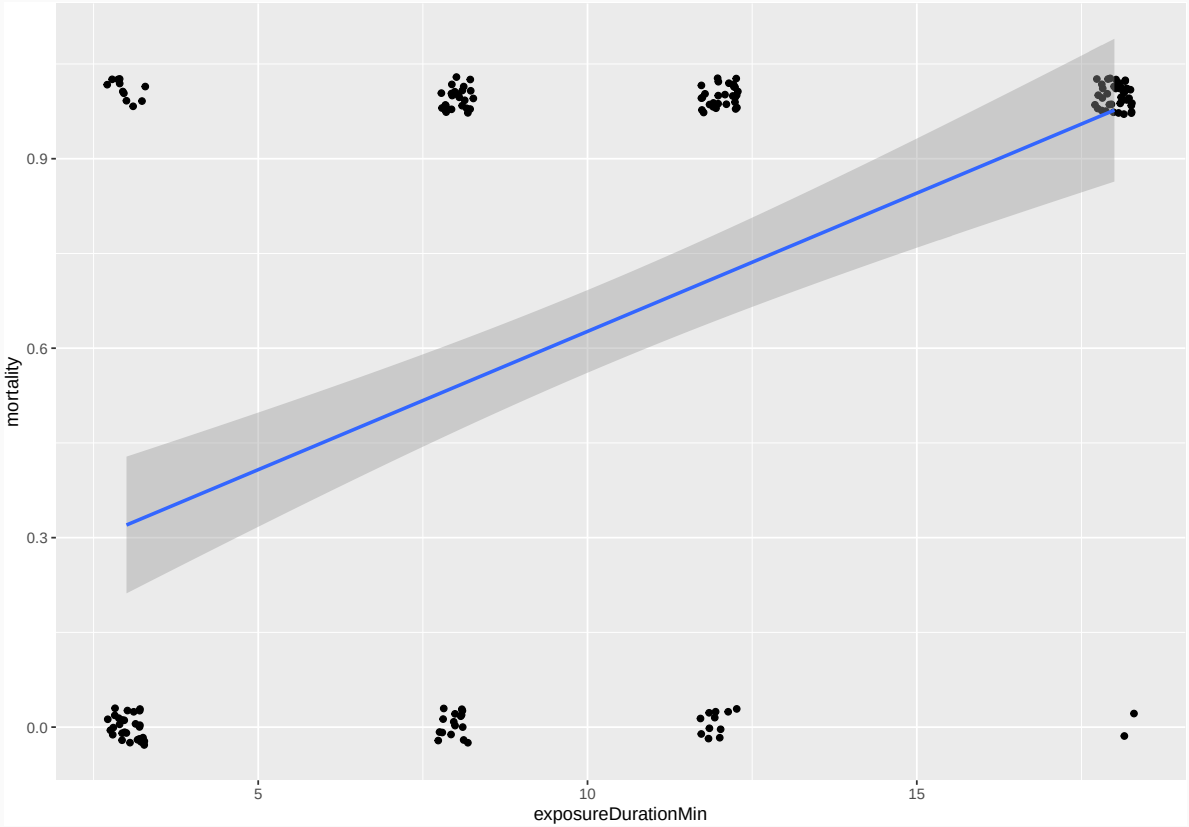
- ¿Podemos aplicar una regresión lineal a estos datos?
- ¿la relación entre X e Y es lineal?

Modelos lineales generalizados

- ¿Podemos aplicar una regresión lineal a estos datos?
- ¿la relación entre X e Y es lineal?
- ¿esperamos que los residuos sean normales?

Modelos lineales generalizados

```
lmgupp <- lm(mortality ~ exposureDurationMin, data = gupp)
ggplot(gupp, aes(x = exposureDurationMin,y = mortality)) +
  geom_point(position = position_jitter(width = .3,height = .03)) +
  geom_smooth(method = "lm")
```



Modelos lineales generalizados

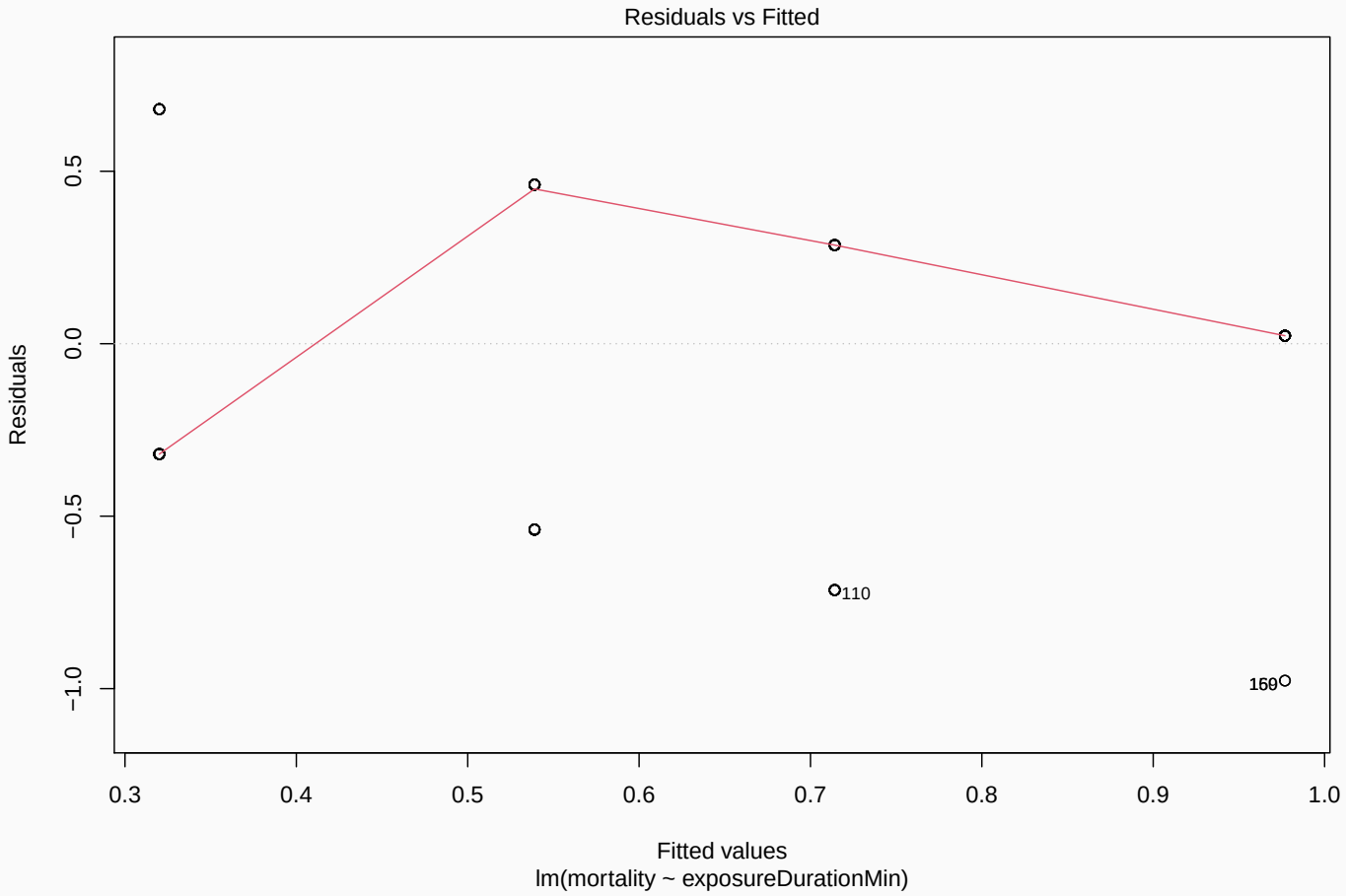
- Para valores muy bajos o muy altos de exposición, la mortalidad es $<$ **0** o $>$ **1**

Modelos lineales generalizados

- Para valores muy bajos o muy altos de exposición, la mortalidad es $< \mathbf{0}$ o $> \mathbf{1}$
- ¿y los residuos?

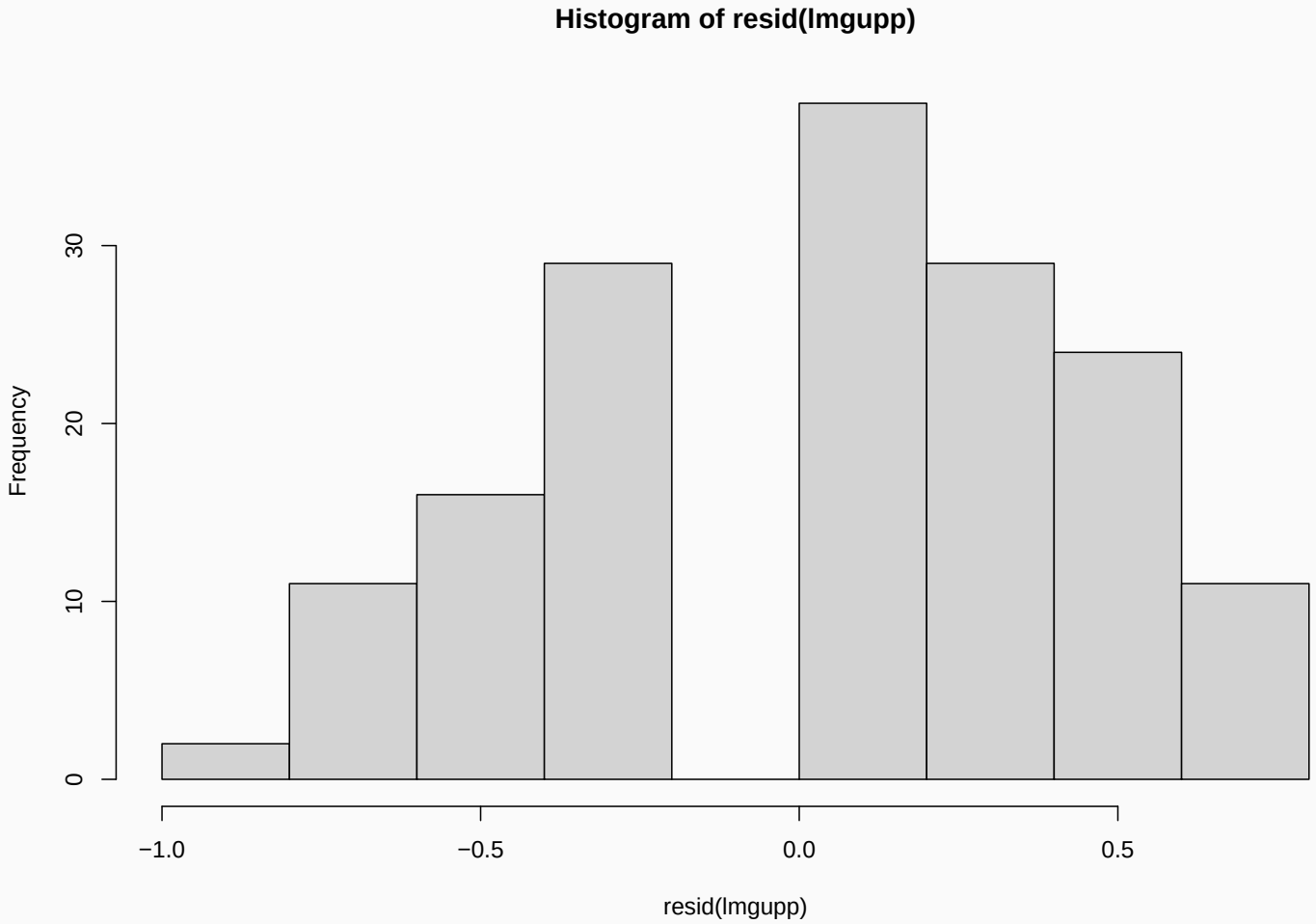
Modelos lineales generalizados

- Para valores muy bajos o muy altos de exposición, la mortalidad es $< \mathbf{0}$ o $> \mathbf{1}$
- ¿y los residuos?



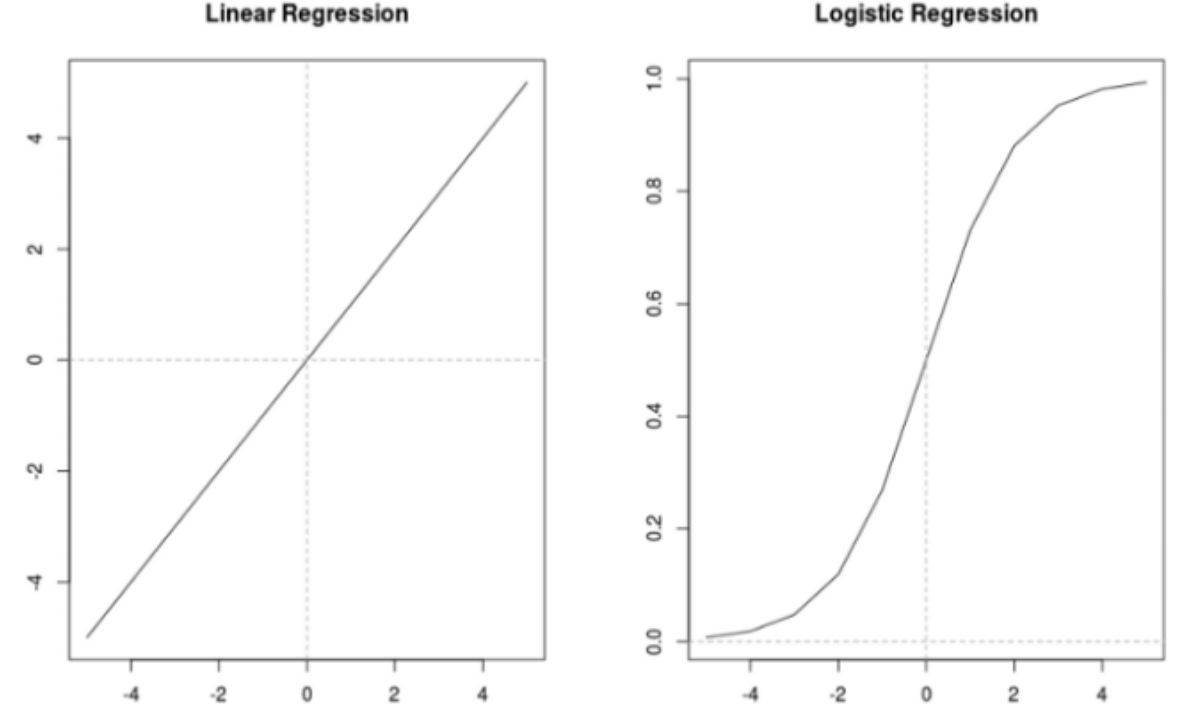
Modelos lineales generalizados

- Para valores muy bajos o muy altos de exposición, la mortalidad es $<$ **0** o $>$ **1**
- ¿y los residuos?

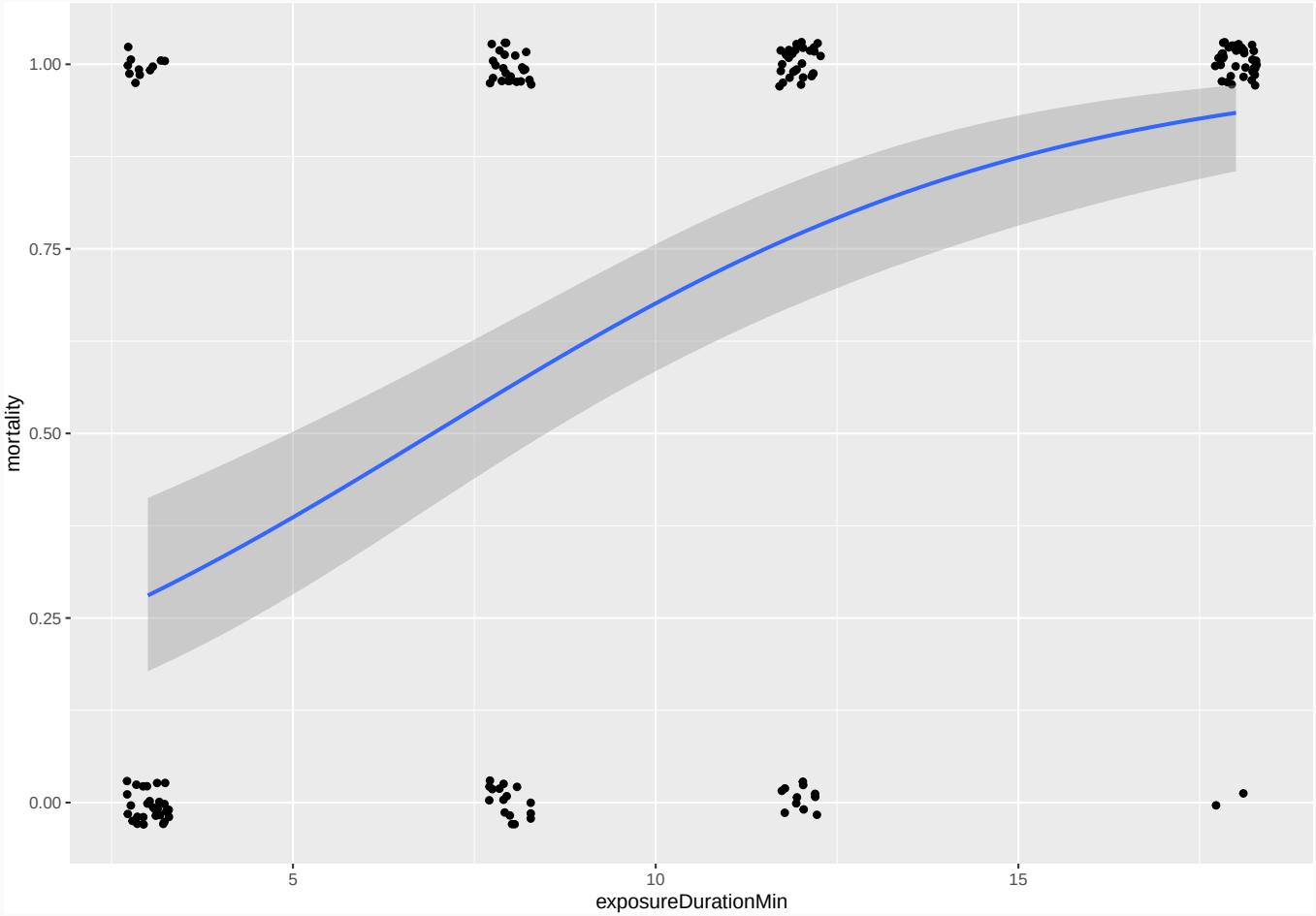


Modelos lineales generalizados

En este caso, queremos modelar la probabilidad de mortalidad en función del tiempo de exposición a temperaturas bajas, con una función limitada entre 0 y 1



Modelos lineales generalizados



Modelos lineales generalizados

Este tipo de modelos, que permiten modelar respuestas *no normales*, se llaman **Modelos lineales generalizados** (*Generalized Linear Models*, GLM).

Tienen tres componentes:

- Distribución estadística de la variable respuesta

Modelos lineales generalizados

Este tipo de modelos, que permiten modelar respuestas *no normales*, se llaman **Modelos lineales generalizados** (*Generalized Linear Models*, GLM).

Tienen tres componentes:

- Distribución estadística de la variable respuesta
- Variables predictoras

Modelos lineales generalizados

Este tipo de modelos, que permiten modelar respuestas *no normales*, se llaman **Modelos lineales generalizados** (*Generalized Linear Models*, GLM).

Tienen tres componentes:

- Distribución estadística de la variable respuesta
- Variables predictoras
- Función de enlace

Modelos lineales generalizados

- En un modelo con variable respuesta **binaria**, la distribución es la distribución **binomial**.

Modelos lineales generalizados

- En un modelo con variable respuesta **binaria**, la distribución es la distribución **binomial**.
- Las variables predictoras son equivalentes a un modelo lineal.

Modelos lineales generalizados

- En un modelo con variable respuesta **binaria**, la distribución es la distribución **binomial**.
- Las variables predictoras son equivalentes a un modelo lineal.
- La función de enlace nos permite modelar nuestra respuesta $a + b \cdot x_i$ en el intervalo $[0, 1]$, en vez de que tome cualquier valor entre $[-\infty, \infty]$

Modelos lineales generalizados

- Función de enlace

Usamos la función logística:

$$Pr(mortalidad_i) = \frac{e^{a+bx_i}}{1 + e^{a+bx_i}}$$

* La función enlace se aplica a la variable respuesta, por lo que reordenamos la ecuación previa:

$$Pr(mortalidad_i) = p_i = g(a + bx_i)$$
$$g^{-1}(p_i) = a + bx_i \quad (1)$$

La función inversa de la logística se llama "logit". Esta, por fin, es nuestra función de enlace:

$$logit(p_i) = a + bx_i$$

De esta manera, para cualquier valor de a, b, x_i, la respuesta estará acotada entre $[0, 1]$.

Modelos lineales generalizados

Función de enlace: Transforma la estimación del modelo para que se ajuste a la distribución de la variable respuesta.

Modelos lineales generalizados

- Ya tenemos todos los ingredientes para ajustar nuestro primer GLM

```
glm1 <- glm(mortality ~ exposureDurationMin,
            data = gupp,
            family = "binomial")
```

que se corresponde con

$$logit(Pr(mortalidad_i))a + bexposure_i)$$

Modelos lineales generalizados

```
summary(glm1)
```

```
##
## Call:
## glm(formula = mortality ~ exposureDurationMin, family = "binomial",
##     data = gupp)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3332  -0.8115   0.3688   0.7206   1.5943
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.66081    0.40651  -4.086 4.40e-05 ***
## exposureDurationMin  0.23971    0.04245   5.646 1.64e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 209.55  on 159  degrees of freedom
## Residual deviance: 164.69  on 158  degrees of freedom
```

```
coef(glm1)
```

```
##          (Intercept) exposureDurationMin
##           -1.6608075            0.2397113
```

Estos coeficientes están en escala logit. Debemos transformarlos a escala natural (probabilidades) para interpretarlos. Para ello, usamos la función logística:

```
plogis(coef(glm1))
```

```
##          (Intercept) exposureDurationMin
##            0.1596536            0.5596425
```

La probabilidad de que un pez muera en condiciones basales es del 0.16

GLM

- distribuciones continuas y discretas

GLM

- distribuciones continuas y discretas
- likelihood (WS p814)

GLM

- distribuciones continuas y discretas
- likelihood (WS p814)
- esquema general: distribución de residuos, fórmula, función de enlace

GLM

- distribuciones continuas y discretas
- likelihood (WS p814)
- esquema general: distribución de residuos, fórmula, función de enlace
- regresión logística (WS p701)

GLM

- distribuciones continuas y discretas
- likelihood (WS p814)
- esquema general: distribución de residuos, fórmula, función de enlace
- regresión logística (WS p701)
- regresión de conteos (poisson, negbin)

GLM

- distribuciones continuas y discretas
- likelihood (WS p814)
- esquema general: distribución de residuos, fórmula, función de enlace
- regresión logística (WS p701)
- regresión de conteos (poisson, negbin)
- selección de modelos (AIC)