Clase 15 Introducción Modelos Lineales Generalizados

Curso Introducción al Análisis de datos con R para la acuicultura.

Dr. José A. Gallardo y Dra. María Angélica Rueda.
jose.gallardo@pucv.cl | Pontificia Universidad Católica de Valparaíso

31 July 2021

PLAN DE LA CLASE

1.- Introducción

-Modelos lineales generalizados ¿Por qué y para qué?

-Componentes de un modelo lineal generalizado (MLG)

-Ecuación del MLG.

-Interpretación de MLG con R.

2.- Práctica con R y Rstudio cloud

-Ajustar modelos lineales generalizados.

-Realizar gráficas avanzadas con ggplot2.

-Elaborar un reporte dinámico en formato pdf.

¿POR QUÉ USAR MODELOS LINEALES GENERALIZADOS?

- 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).

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. La transformación se realiza para producir aproximadamente:

-Normalidad

-Homogeneidad

-Linealidad

INTRODUCCIÓN

- 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.

Alternativa: **Modelos Lineales Generalizados (MLG)** (Nelder y Wedderburn, 1972)

MODELOS LINEALES GENERALIZADOS

Los modelos lineales generalizados 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.**
- Modelo de regresión logística.
- Modelos log-lineales: **para tablas de contingencia**.

COMPONENTES DE UN MODELO LINEAL GENERALIZADO

1. Componente aleatorio:

La variable respuesta y su distribución de probabilidad: ***la familia exponencial natural***.

Por ej.

$$Y_i \sim N(\mu_i, \sigma^2)$$

$$E(Y_i) = \mu_i$$

$$var(Y_i) = \sigma^2$$

TIPOS DE DISTRIBUCIONES

La familia exponencial natural contiene las siguientes distribuciones:

- Normal, Poisson, Binomial (Binaria: 0 y 1, caso particular de la binomial), Gamma, Binomial Negativa, Multinomial, entre otras.
- La elección del tipo de distribución a usar debe realizarse ***a priori*** por el analista de los datos, dependerá de la naturaleza de la variable respuesta (como se generaron los datos).
- Cada distribución se caracteriza por su relación **media** (parámetro de posición) y **varianza** (parámetro de dispersión).

FORMA EN QUE SE GENERAN LOS DATOS

Hay que examinar cuidadosamente los datos, principalmente en cuantos a asimetría, naturaleza continua o discreta e intervalo de variación.

Distribución	Origen
Normal	Simetría y el intervalo de variación es la recta de los reales.
Poisson	Conteos o datos continuos con varianza similar a la media.
Binomial	Datos en forma de proporciones (nº de éxitos respecto a un total).
Bernoulli	Caso especial de Binomial. Toma solo valores de 0 y 1.
Gama	Datos continuos asimétricos y con valores positivos, coeficiente de variación constante.
Binomial Negativa	Número de experimentos de Bernoulli hasta la consecución del k-ésimo éxito.

RELACIÓN ENTRE EL PARÁMETRO DE POSICIÓN Y EL DE DISPERSIÓN

Distribución	Posición	Dispersión
Normal	$E(X) = \mu$	$var(Y) = \sigma^2$
Poisson	$E(X) = \mu$	$var(Y) = \mu$
Binomial	$E(X) = np$	$var(Y) = np(1-p)$
Bernoulli	$E(X) = p$	$var(Y) = p(1-p)$
Binomial Negativa	$E(X) = \mu$	$var(Y) = \mu + \mu^2/k$

COMPONENTES DE UN MODELO LINEAL GENERALIZADO

2. Componente sistemático: ***(es lineal e identifica la(s) variables predictoras)***

- Las variables predictoras en el modelo pueden ser continuas, categóricas, funciones polinomiales, interacciones.
- Relaciona un vector η, (llamado ***predictor lineal***) con las variables predictoras X a través de un modelo lineal, esto es:

$$\eta(X_{i1}, ..., X_{ip}) = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}$$

COMPONENTES DE UN MODELO LINEAL GENERALIZADO

3. Función de enlace:

Conecta los componentes ***aleatorio*** y ***sistemático***. Relaciona el valor esperado de la variable aleatoria con el ***predictor lineal*** mediante

$$g(\mu_i) = \eta(X_{i1}, ..., X_{ip})$$

Cada distribución posee su función de enlace, hay distribuciones que tienen más de una (por ej. Binomial (enlace **logit**, **probit** o **complemento log-log**).

EL MODELO LINEAL VISTO COMO UN MODELO LINEAL GENERALIZADO

El modelo lineal clásico es un caso particular de modelo lineal generalizado

- **Componente aleatorio:** Las Y_i son variables aleatorias independientes

$$Y_i \sim N(\mu_i, \sigma^2)$$

- **Componente sistemático:** El predictor lineal es

$$\eta(X_{i1}, ..., X_{ip}) = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}$$

- **Función de enlace: Identidad**

$$g(\mu_i) = \mu_i$$

DISTRIBUCIONES Y SUS FUNCIONES DE ENLACE

Distribución	Enlace (canónico en negrita)
Normal	**Identidad**, log, inverso.
Poisson	**Log**, identidad, raíz cuadrada.
Binomial	**Logit**, probit, complemento log-log.
Binomial Negativa	**Log**, identidad, raíz cuadrada.

¿QUÉ MODELOS COMPARAR?

- **Modelo nulo:** No ofrece ninguna explicación para los datos, se expresa como $y \sim 1$ es el modelo más simple, tiene solo un parámetro, representa la media global μ_i para todos los y. Toda la variación de y se le atribuye al **componente aleatorio**.
- **Modelo saturado:** Es el modelo más extremo, tiene n parámetros.
- **Modelo corriente:** es el que intentamos buscar, aquel que explique la mayor parte de la variación de los datos, pero que use el menor número de parámetros posibles.

¿CÓMO CONOZCO EL AJUSTE DEL MODELO?

Una medida de bondad de ajuste del modelo es la **Deviance** (también llamada **devianza**).

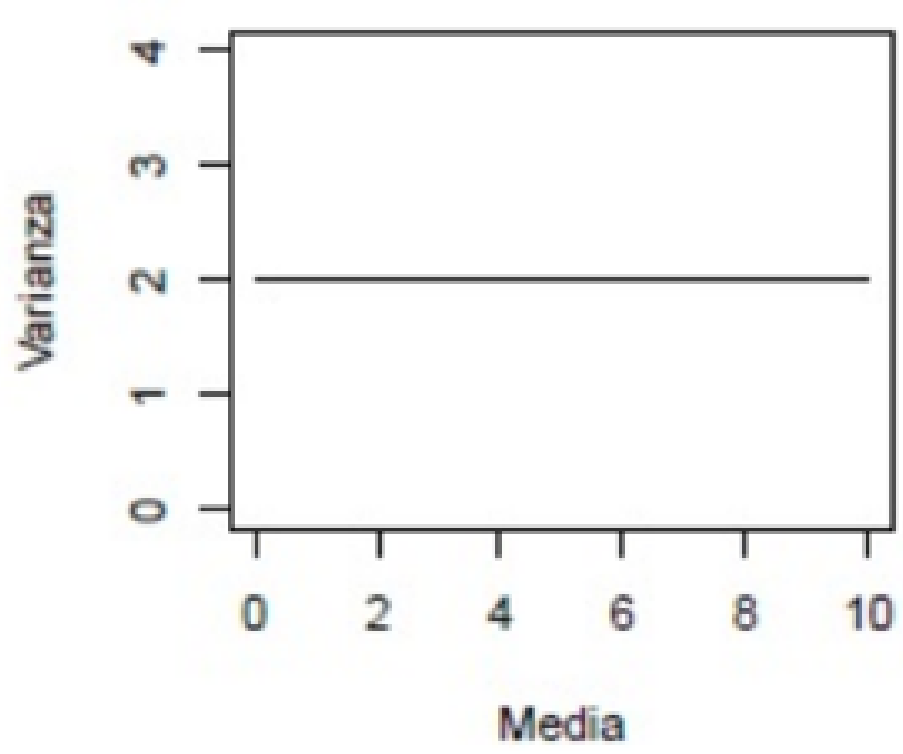
$$Dev = -2log(\frac{L(Modelo_{corriente})}{L(Modelo_{saturado})}$$

$$Dev \sim \chi^2$$

- **Modelo nulo**: tiene la **máxima devianza**.
- **Modelo saturado**: Tiene **devianza igual a cero**.
- **Modelo corriente**: es el que deja una devianza residual **lo más pequeña posible**.

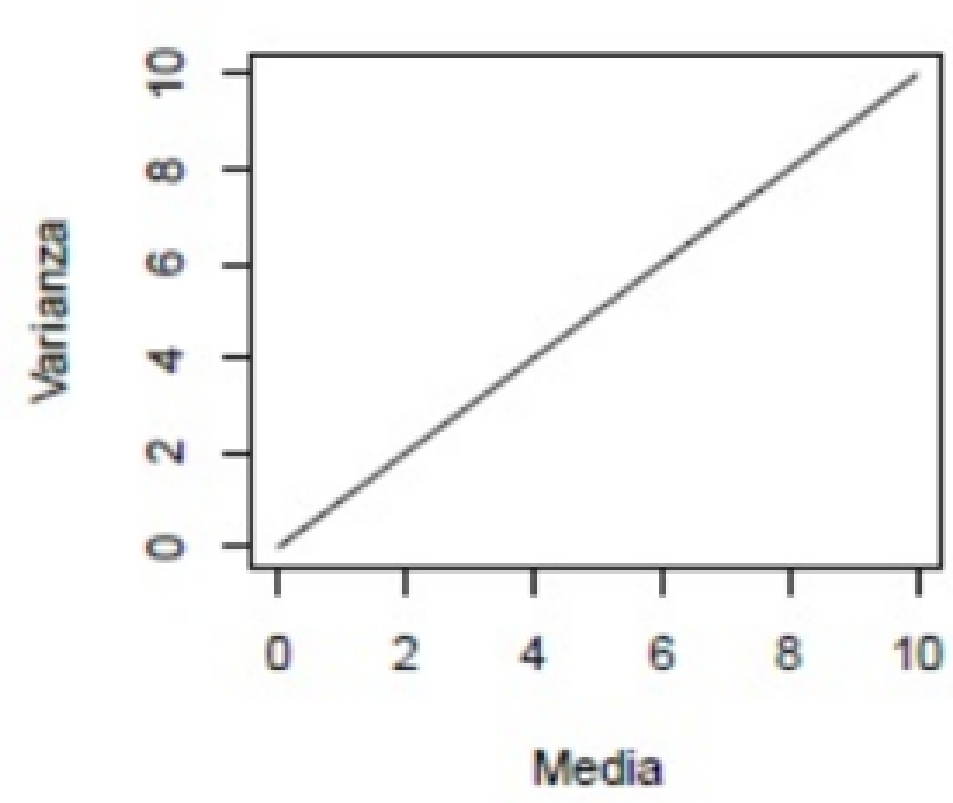
RELACIÓN MEDIA-VARIANZA PARA IDENTIFICAR LA POSIBLE DISTRIBUCIÓN DE MI VARIABLE RESPUESTA

El supuesto central que se hace en los modelos lineales es que la varianza es constante, así que al variar la media la varianza se mantiene constante.



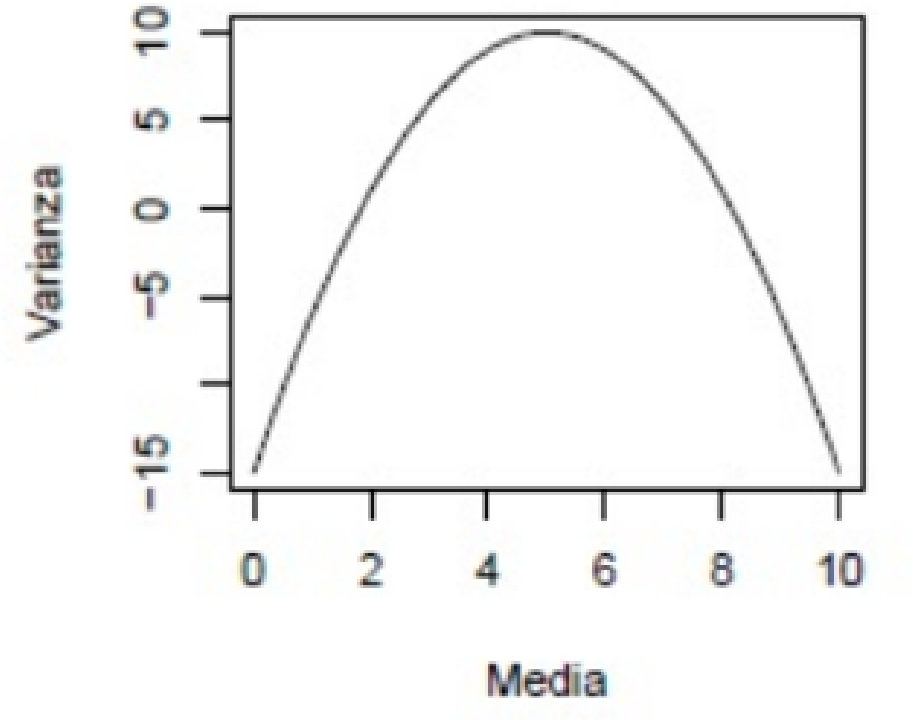
RELACIÓN MEDIA-VARIANZA PARA IDENTIFICAR LA POSIBLE DISTRIBUCIÓN DE MI VARIABLE RESPUESTA

En el caso de variables respuestas de conteo expresadas como números enteros y en donde puede haber muchos ceros en los datos, la varianza se suele incrementar linealmente con la media.



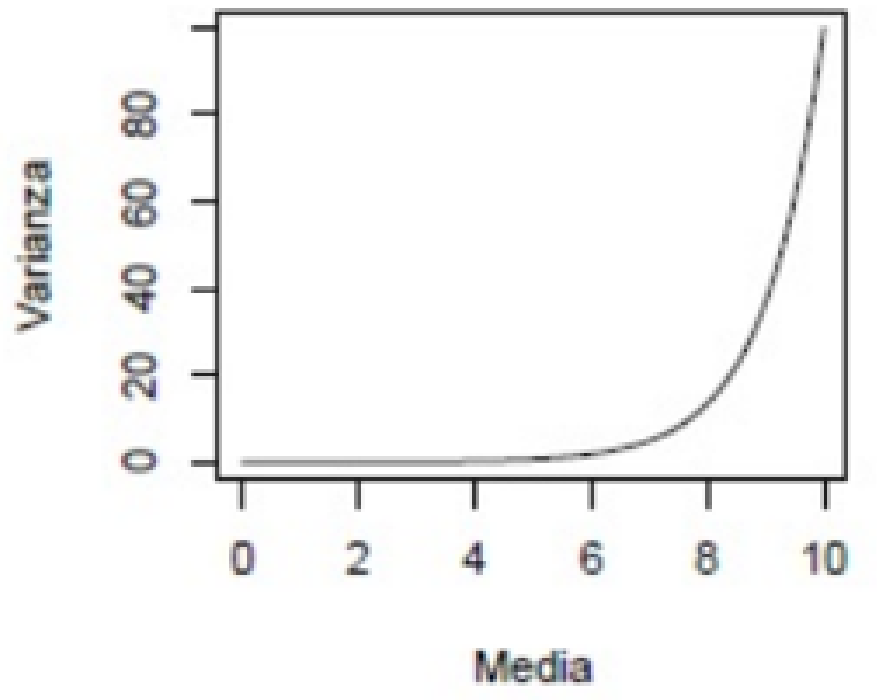
RELACIÓN MEDIA-VARIANZA PARA IDENTIFICAR LA POSIBLE DISTRIBUCIÓN DE MI VARIABLE RESPUESTA

Cuando la variable respuesta sea proporciones de eventos es muy posible que la varianza se comporte en forma de U invertida.



RELACIÓN MEDIA-VARIANZA PARA IDENTIFICAR LA POSIBLE DISTRIBUCIÓN DE MI VARIABLE RESPUESTA

Cuando la variable respuesta se aproxime a una distribución gamma, entonces la varianza se incrementa de una manera no lineal con respecto a la media.



CONCEPTOS CLAVE

- Elegimos a priori una distribución para la variable respuesta, basada en su naturaleza y su relación media-varianza.
- Elijo una función de enlace que proyecta la predicción en el espacio de la variable respuesta.
- Construyo un modelo con variables X predictoras.
- Obtengo la **Devianza** del modelo que me interesa y la comparo con la del modelo saturado.
- El modelo debe tener sentido desde el punto de vista particular de la aplicación.

LIBRERÍA PARA AJUSTAR MODELOS LINEALES GENERALIZADOS

```
library(stats)
```

Función **glm()**

EJEMPLO: SALMÓN DEL ATLÁNTICO

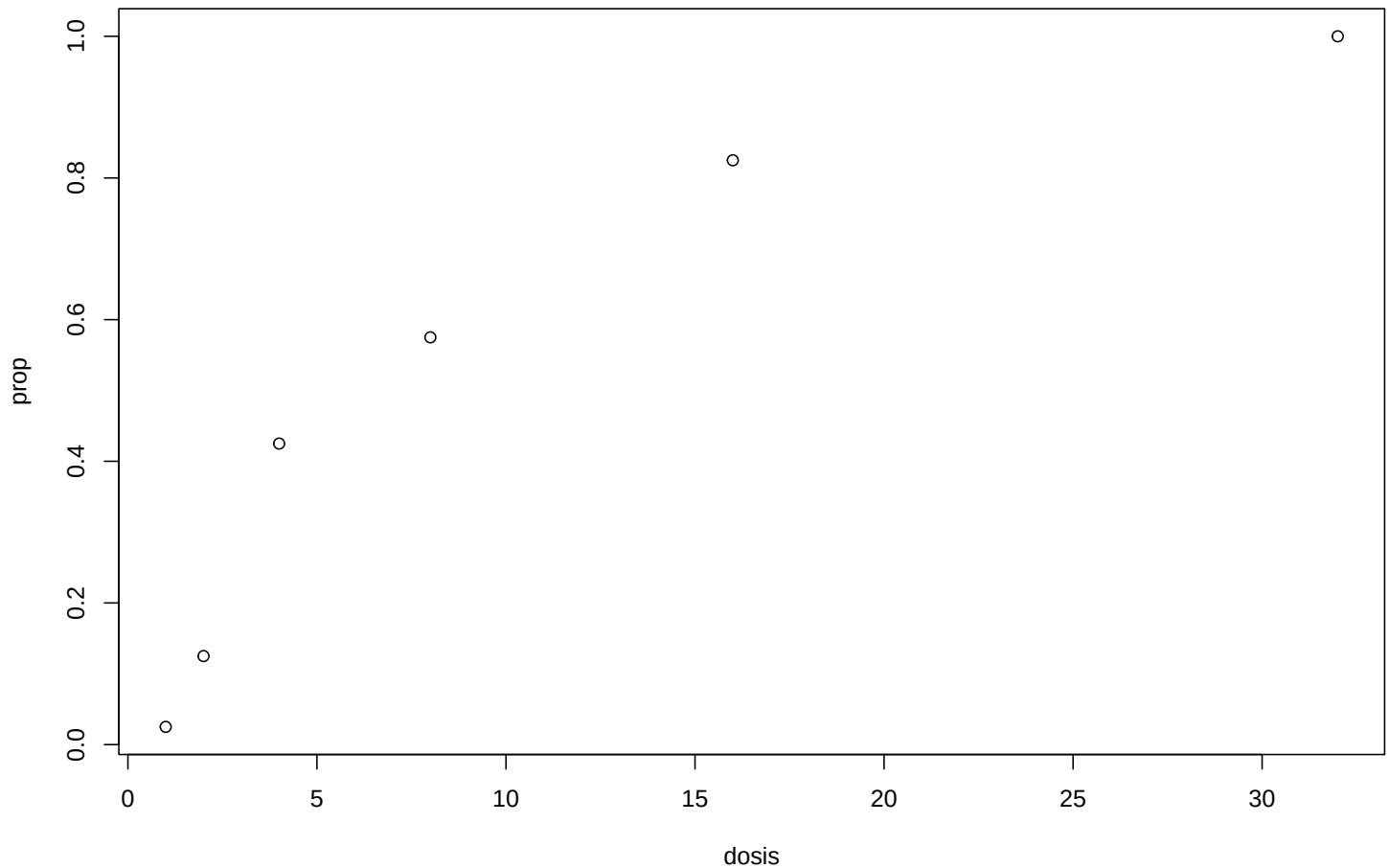
Con el objeto de estudiar el efecto del tratamiento veterinario sobre la mortalidad del salmón del Atlántico, se consideraron seis grupos de 40 salmones, sometiendo cada grupo a una dosis diferente del tratamiento veterinario, y se reportaba el número de muertos en cada grupo.

Table 1: Tabla de datos

dosis	muertos	no muertos	total individuos
1	1	39	40
2	5	35	40
4	17	23	40
8	23	17	40
16	33	7	40
32	40	4	40

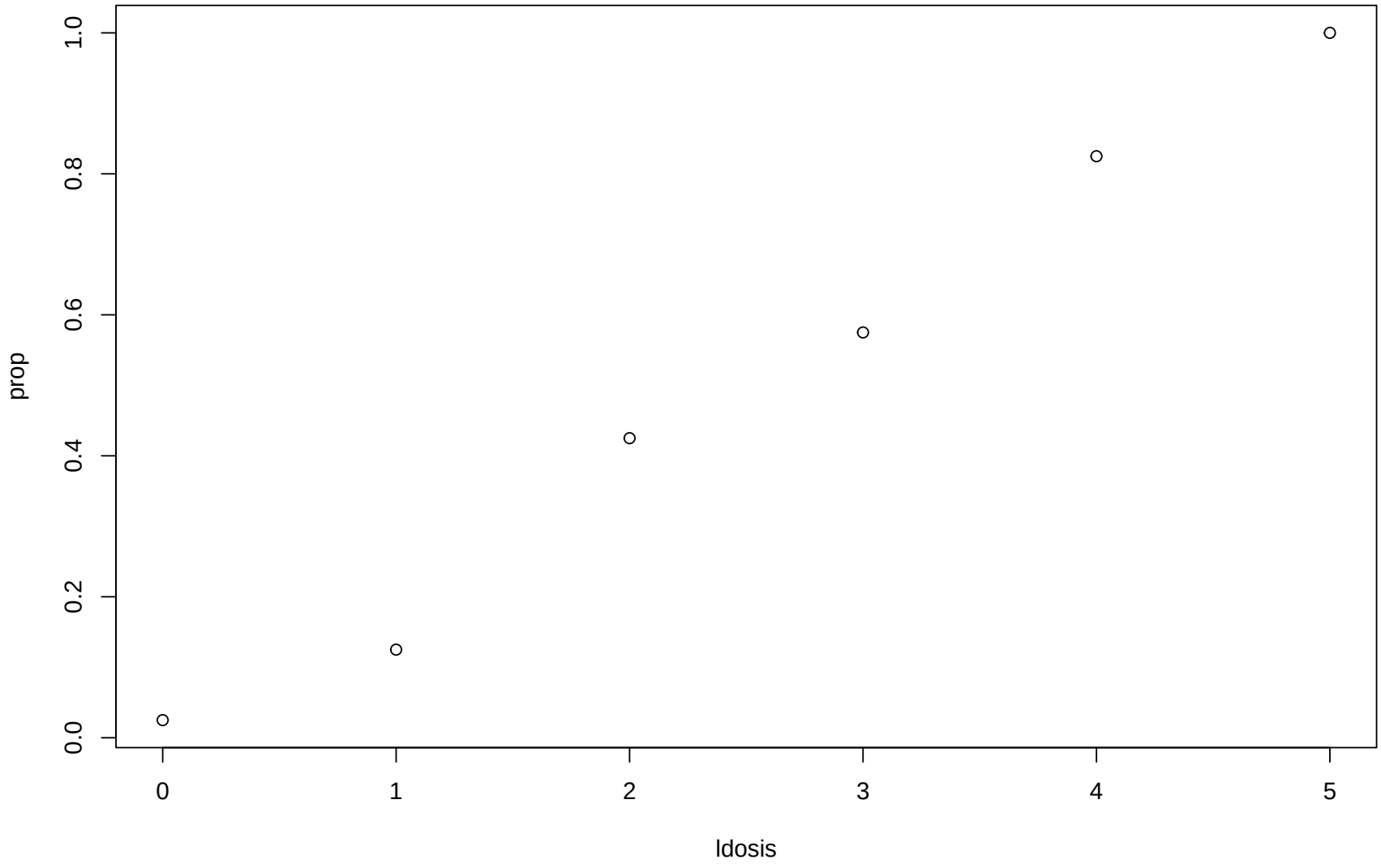
PROPORCIÓN DE SALMONES MUERTOS

```
muertos <- salmones$muertos
total <- salmones$`total individuos`
dosis <- salmones$dosis
prop<-muertos/total
plot(dosis,prop)
```



PROPORCIÓN DE SALMONES MUERTOS

```
ldosis<-log(dosis,2)
plot(ldosis,prop)
```



EJEMPLO: SALMÓN DEL ATLÁNTICO

Se debe generar la variable *y* para realizar el ajuste del modelo binomial con la función **glm()**, y debe tener la siguiente estructura:

```
muertos <- salmones$muertos
total <- salmones$`total individuos`
y<- cbind(muertos,total-muertos)
```

EJEMPLO: SALMÓN DEL ATLÁNTICO

- Modelo ajustado con un enlace **logit**:

```
ajustelogit<-glm(y~ldosis,family=binomial(link="logit"))
plogit<-1-pchisq(6.313,4)
```

```
Call:  glm(formula = y ~ ldosis, family = binomial(link = "logit"))

Coefficients:
(Intercept)       ldosis
     -3.204        1.269

Degrees of Freedom: 5 Total (i.e. Null);  4 Residual
Null Deviance:            147
Residual Deviance: 6.313    AIC: 27.51

Call:
glm(formula = y ~ ldosis, family = binomial(link = "logit"))

Deviance Residuals:
       1         2         3         4         5         6
-0.48867  -0.02108   1.12802  -0.92598  -0.74044   1.84273

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.2044     0.4207  -7.616 2.62e-14 ***
ldosis        1.2685     0.1505   8.431  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 147.005  on 5  degrees of freedom
Residual deviance:   6.313  on 4  degrees of freedom
AIC: 27.513

Number of Fisher Scoring iterations: 4
```

EJEMPLO: SALMÓN DEL ATLÁNTICO

- Modelo ajustado con un enlace **probit**:

```
ajusteprobit<-glm(y~ldosis,family=binomial(link="probit"))
pprobit<-1-pchisq(4.871,4)
```

```
Call:  glm(formula = y ~ ldosis, family = binomial(link = "probit"))

Coefficients:
(Intercept)       ldosis
    -1.8791       0.7456

Degrees of Freedom: 5 Total (i.e. Null);  4 Residual
Null Deviance:           147
Residual Deviance: 4.871    AIC: 26.07

Call:
glm(formula = y ~ ldosis, family = binomial(link = "probit"))

Deviance Residuals:
       1         2         3         4         5         6
-0.19490  -0.06661   0.99304  -0.84360  -0.71466   1.61867

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.87910    0.22129  -8.492   <2e-16 ***
ldosis       0.74564    0.07774   9.591   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 147.005  on 5  degrees of freedom
Residual deviance:   4.871  on 4  degrees of freedom
AIC: 26.071

Number of Fisher Scoring iterations: 4
```

EJEMPLO: SALMÓN DEL ATLÁNTICO

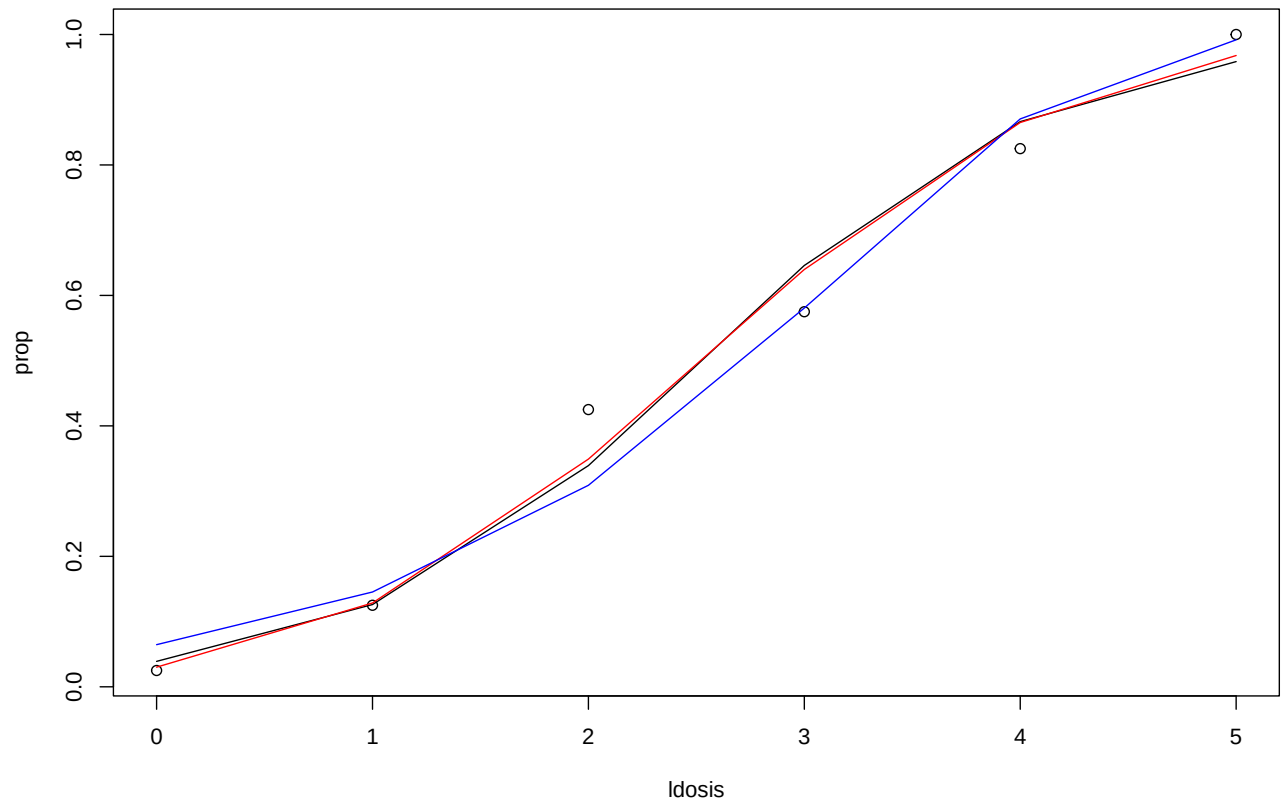
- Modelo ajustado con un enlace **complemento log-log**:

```
ajusteclog<-glm(y~ldosis,family=binomial(link="cloglog"))
pclog<-1-pchisq(5.1964,4)
```

```
Call:
glm(formula = y ~ ldosis, family = binomial(link = "cloglog"))

Deviance Residuals:
      1          2          3          4          5          6
-1.15540  -0.37278   1.54541  -0.07278  -0.82095   0.80931

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.70649    0.31488  -8.595   <2e-16 ***
ldosis       0.85541    0.09408   9.093   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 147.0050  on 5  degrees of freedom
Residual deviance:   5.1964  on 4  degrees of freedom
AIC: 26.396

Number of Fisher Scoring iterations: 5
```

EJEMPLO: SALMÓN DEL ATLÁNTICO

```
library(boot)
plot(ldosis,prop)
lines(ldosis, fitted.values(ajustelogit), type="l")
lines(ldosis, fitted.values(ajusteprobit), type="l",
      col="red")
lines(ldosis, fitted.values(ajusteclog), type="l",
      col="blue")
```



EJEMPLO: SALMÓN DEL ATLÁNTICO

TABLA RESUMEN

Link	P valores	AIC
Logit	0.1769609	27.513
Probit	0.3007916	**26.071**
Clog-log	0.2677327	26.396

- Con ninguna de las tres funciones de enlace se rechaza la adecuación del modelo al 5%.
- Notemos que usando el enlace **probit** el criterio de AIC dio menor y además se observa en el gráfico que la curva ajustada usando dicha función de enlace ajusta mejor los valores observados, por lo tanto será el ajuste seleccionado.

RESUMEN DE LA CLASE

1). Revisión de conceptos: modelos lineales generales, modelos lineales generalizados.

2). Construir y ajustar modelos lineales generalizados.