Modelos lineales generalizados

g(.) = Función liga o conectora

- Especifica la función para E (y) = μ
- Conecta los componentes aleatorio y sistemático

$$g(E(y)) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

Componente Aleatorio

Componente Sistemático

- Identifica a la variable respuesta y
- Asume una distribución de probabilidad para y

- Especifica las variables explicativas del modelo como una combinación lineal $\{x_j\}$. Es decir, las variables explicativas entran linealmente como predictores a la derecha de la ecuación del modelo.
- El subíndice para cada una de las x es para enfatizar que estas variables x son fijas (no aleatorias)

Así, por ejemplo, los GLM´s, permiten especificar si la variable respuesta es continua (y ~Normal), si es categórica con dos posibles respuestas en un número determinado de intentos (y ~Binomial), o si esta corresponde a conteos, y por lo tanto, sólo toma valores de enteros no negativos (y ~Poisson).

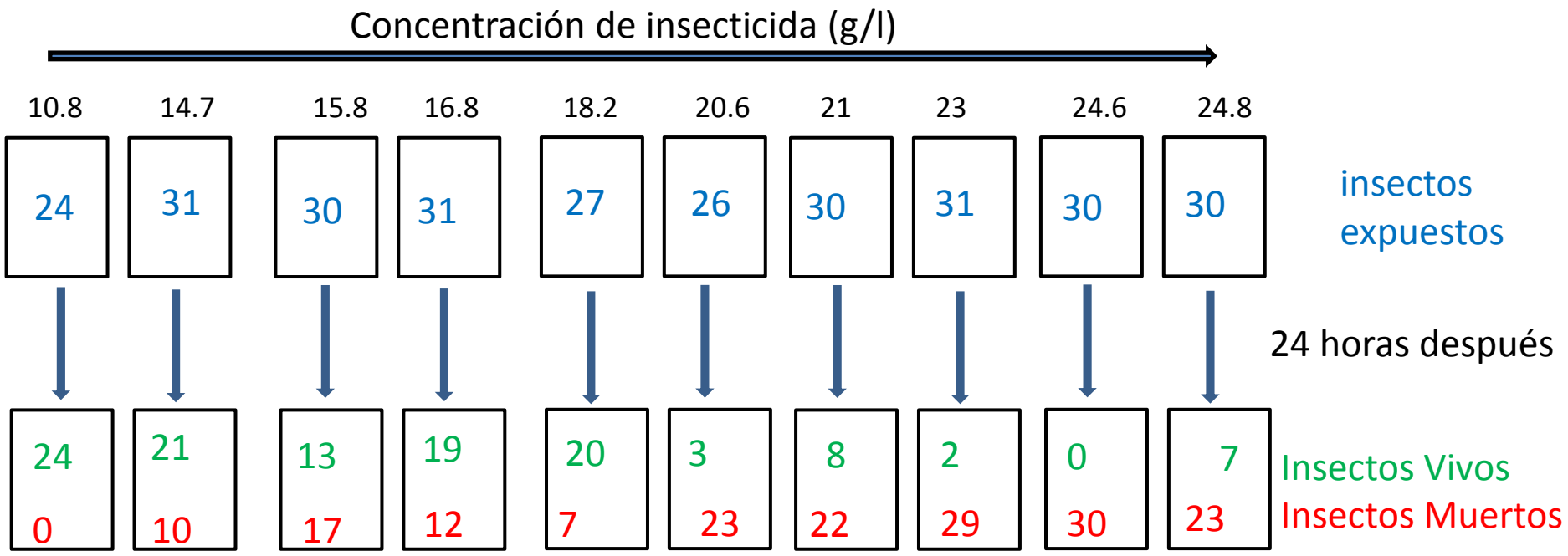
Y además, las variables que compondrán el componente sistemático del modelo, no sólo serán categóricas o continuas, sino que podrán ser de ambos tipos

Table 3.5. Types of Generalized Linear Models for Statistical Analysis

Random Component	Link	Systematic Component	Model
Normal	Identity	Continuous	Regression
Normal	Identity	Categorical	Analysis of variance
Normal	Identity	Mixed	Analysis of covariance
Binomial	Logit	Mixed	Logistic regression
Multinomial	Logits	Mixed	Multinomial response
Poisson	Log	Mixed	Loglinear

Tomado de: Agresti, A. 2007. An Introduction to Categorical Data Analysis. John Wiley & Sons. USA

```
> insecto <- read.csv("C:/Users/Karla/Desktop/Categoricos/insecto.csv")
> variable.names(insecto)
[1] "concentracion" "afectados"     "expuestos"
> prop_muertos<-insecto$afectados/insecto$expuestos
> vivos<-insecto$expuestos-insecto$afectados
> insecto1<-data.frame(concentracion, afectados, expuestos, prop_muertos, vivos )
> insecto1
   concentracion afectados expuestos prop_muertos vivos
1           24.8        23        30    0.7666667     7
2           24.6        30        30    1.0000000     0
3           23.0        29        31    0.9354839     2
4           21.0        22        30    0.7333333     8
5           20.6        23        26    0.8846154     3
6           18.2         7        27    0.2592593    20
7           16.8        12        31    0.3870968    19
8           15.8        17        30    0.5666667    13
9           14.7        10        31    0.3225806    21
10          10.8         0        24    0.0000000    24
```



Family	Link	Mean Function
gaussian	identity	$\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$
binomial	logit	$\mu_i = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})}$
binomial	probit	$\mu_i = \Phi(\mathbf{x}_i'\boldsymbol{\beta})$
binomial	cloglog	$\mu_i = 1 - \exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))$
poisson	log	$\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$
poisson	identity	$\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$
poisson	sqrt	$\mu_i = (\mathbf{x}_i'\boldsymbol{\beta})^2$
gamma	inverse	$\mu_i = (\mathbf{x}_i'\boldsymbol{\beta})^{-1}$
gamma	identity	$\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$
gamma	log	$\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$
inverse gaussian	inverse squared	$\mu_i = (\mathbf{x}_i'\boldsymbol{\beta})^{-1/2}$

	logit	**probit**	**complementary log- log**
$\eta = g(\theta)$	$\ln\frac{\theta}{1-\theta}$	$\Phi^{-1}(\theta)$	$\ln[-\ln(1-\theta)]$
$\theta = g^{-1}(\eta)$	$\frac{\exp(\eta)}{1+\exp(\eta)}$	$\Phi(\eta)$	$1-\exp[-\exp(\eta)]$
$\frac{\partial\eta}{\partial\mu}$	$\frac{\theta}{\theta(1-\theta)}$	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y^2\right)$	$-\frac{1}{(1-\theta)\ln(1-\theta)}$
w_{ii}	$n\theta(1=\theta)$	$\frac{n}{\theta(1=\theta)}\left(\frac{\sqrt{2\pi}}{\exp\left(-\frac{1}{2}y^2\right)}\right)^2$	$\frac{n(1-\theta)}{\theta}\ln^2(1-\theta)$

Note: ln =natural logarithm, Φ =cumulative distribution of the standardized normal.

Usando liga Logit

$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 * x$

p= probabilidad de éxito

ln= logaritmo natural

Usando liga probit

$\overline{\Phi^{-1}}(p) = \beta_0 + \beta_1 * x$

$\overline{\Phi^{-1}}$= Inversa de distribución normal. Lo que se obtiene al sustituir las betas y el valor de *x* en la ecuación es un z-score

Usando liga c log log (log log complementaria

$ln(-ln(1-p)) = \beta_0 + \beta_1 * x$

Modelo logit

```
> modeloLogit<-glm(cbind(afectados,vivos)~concentracion,binomial(link="logit"), data=insecto1)
> summary(modeloLogit)

Call:
glm(formula = cbind(afectados, vivos) ~ concentracion, family = binomial(link = "logit"),
    data = insecto1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0670  -1.7117   0.1495   1.7340   2.4150

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.01047    0.77937  -7.712 1.24e-14 ***
concentracion  0.34127    0.04153   8.217  < 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: 138.001  on 9  degrees of freedom
Residual deviance:  37.697  on 8  degrees of freedom
AIC: 69.27

Number of Fisher Scoring iterations: 5
```

$$\ln \frac{probabilidad\ de\ morir}{probabilidad\ de\ vivir} = -6.01047 + 0.34127 \ *\text{concentración}$$

$$\ln \frac{probabilidad\ de\ morir}{probabilidad\ de\ vivir} = -6.01047 + 0.34127 * \text{concentración}$$

$$\ln \frac{probabilidad\ de\ morir}{probabilidad\ de\ vivir} = -6.01047 + 0.34127 * 0$$

Interpretación de β_0

$$e^{-6.01047} = 0.000245 = \frac{probabilidad\ de\ morir}{probabilidad\ de\ vivir}$$

Cuando no hay toxina en el ensayo (concentración de 0), los *odds* de morir (respecto a vivir) son ~ 0. La probabilidad de morir representa un 0.02% de la probabilidad de vivir a esta dosis.

Interpretación de β_1

$$e^{0.34127} = 1.40673$$

Por cada unidad de incremento en la concentración de insecticida, la probabilidad de morir respecto a la probabilidad de vivir aumenta en 1.4 veces. Dicho de otra forma, por cada unidad de incremento en la concentración, los *odds* o posibilidades de morir incrementan en 1.4 veces

Es decir, si a una concentración x, el odds resulta 2, cuando la concentración es x+1, el odds es 2.8 (porque 1.4*2 =2.8), y a una concentración de x+2, el odds es 3.92 (porque 1.4*2.8 = 3.92)

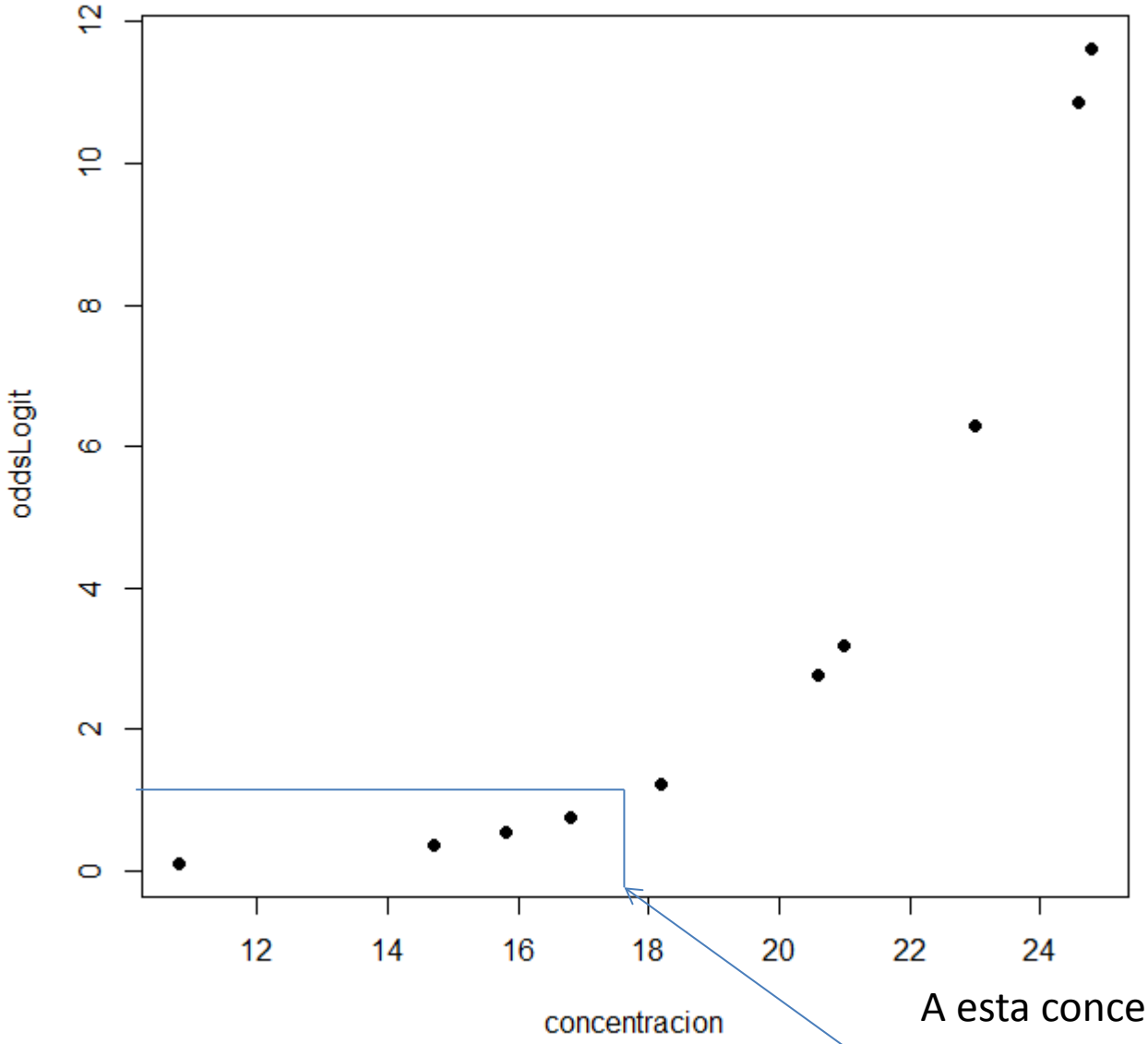
```
> prob.logit<-ilogit(modeloLogit$coef[1] + modeloLogit$coef[2]*concentracion)
> q.logit<-1-prob.logit
> oddsLogit<-prob.logit/q.logit
> cbind(concentracion,prop_muertos, prob.logit,q.logit, oddsLogit)
      concentracion prop_muertos prob.logit    q.logit   oddsLogit
 [1,]          24.8    0.7666667 0.92078190 0.07921810 11.62337797
 [2,]          24.6    1.0000000 0.91565812 0.08434188 10.85650536
 [3,]          23.0    0.9354839 0.86279879 0.13720121  6.28856529
 [4,]          21.0    0.7333333 0.76064038 0.23935962  3.17781404
 [5,]          20.6    0.8846154 0.73491137 0.26508863  2.77232320
 [6,]          18.2    0.2592593 0.54999200 0.45000800  1.22218274
 [7,]          16.8    0.3870968 0.43115528 0.56884472  0.75794899
 [8,]          15.8    0.5666667 0.35014341 0.64985659  0.53880105
 [9,]          14.7    0.3225806 0.27016106 0.72983894  0.37016532
[10,]          10.8    0.0000000 0.08909321 0.91090679  0.09780717
```

```
> prob.logit16.8<-ilogit(modeloLogit$coef[1] + modeloLogit$coef[2]*16.8)
> q.logit16.8<-1-prob.logit16.8
> prob.logit15.8<-ilogit(modeloLogit$coef[1] + modeloLogit$coef[2]*15.8)
> q.logit15.8<-1-prob.logit15.8
> odds16.8.15.8<-(prob.logit16.8/q.logit16.8) / (prob.logit15.8/q.logit15.8)
> odds16.8.15.8
(Intercept)
   1.406733
```

Aquí, al pasar de una dosis de 15.8 a 16.8, el incremento en el *odds* es 1.4

```
> prob.logit11<-ilogit(modeloLogit$coef[1] + modeloLogit$coef[2]*11)
> q.logit11<-1-prob.logit11
> prob.logit10<-ilogit(modeloLogit$coef[1] + modeloLogit$coef[2]*10)
> q.logit10<-1-prob.logit10
> odds11.10<-(prob.logit11/q.logit11) / (prob.logit10/q.logit10)
> odds11.10
(Intercept)
   1.406733
```

Aquí, al pasar de una dosis de 10 a 11, el incremento en el *odds* es 1.4




```
> plot(concentracion,oddsLogit, pch=16)
```

Podemos apreciar que al incrementarse la concentración de insecticida en una unidad, los *odds* de morir respecto a no morir, incrementan 1.4 unidades (respecto a los *odds* en la concentración anterior)

A esta concentración, las probabilidades de morir y de no morir son las mismas (*odds* Logit= 1).
A esa concentración se le llama LD_{50}

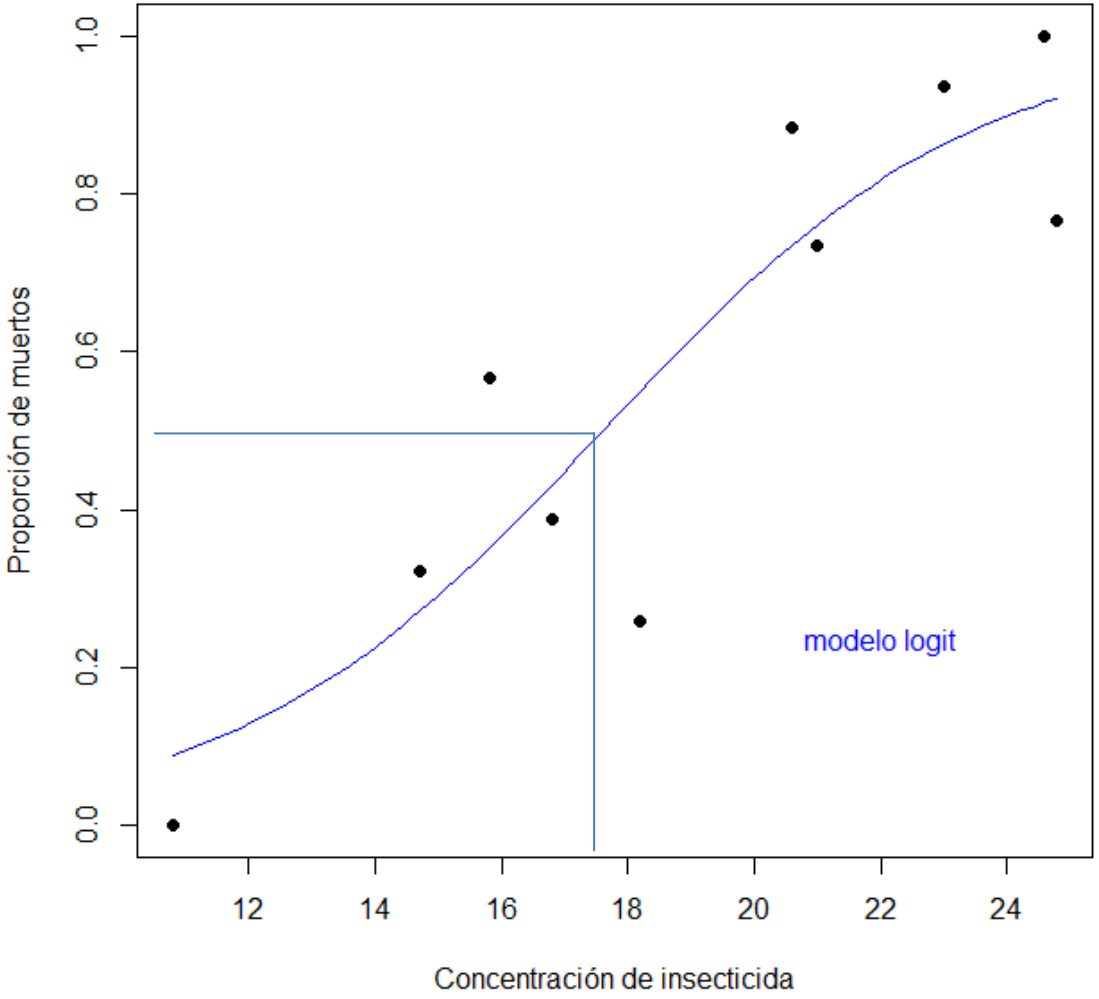
LD_{50} = Dosis Letal media

$$\ln \frac{0.5}{0.5} = -6.01047 + 0.34127 \ *\text{concentración}$$

$$0 = -6.01047 + 0.34127 \ *\text{concentración}$$

$$\frac{6.01047}{0.34127} = \text{concentración}$$

```
> (0-modeloLogit$coef[1])/modeloLogit$coef[2]
(Intercept)
   17.61208
```



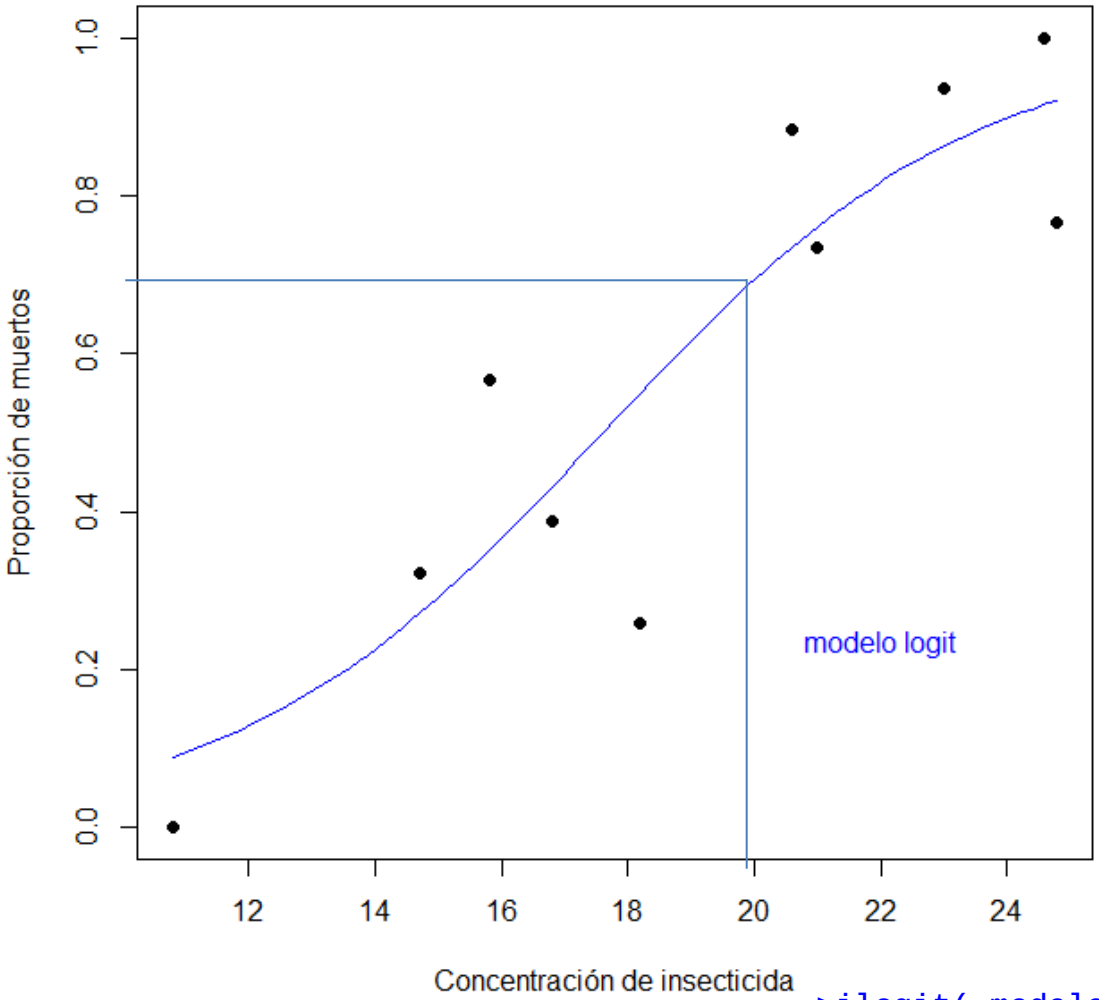
La LD_{50} es 17.61208. A esa concentración, la probabilidad de morir es 0.5 (y la probabilidad de no morir, también es 0.5),y espero que por cada insecto vivo que encuentre, encuentre un insecto muerto.

Dosis Letal 50 (LD_{50}) = Dosis a la que la mitad de los individuos de una población mueren.

```
>plot(insecto1$concentracion,prop_muertos, xlab="Concentración de insecticida", ylab="Proporción de muertos", pch=16, col="black")
> curve(ilogit(modeloLogit$coef[1] + modeloLogit$coef[2]*x),add=T, lty=1,lwd=1.7, col="blue")
> text(22,0.2,labels=c("modelo logit"), pos=3, col=c(4))
```

La probabilidad de morir a una concentracion de 20

$$\hat{p} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 * \text{concentración})}{1+\exp(\hat{\beta}_0 + \hat{\beta}_1 * \text{concentración})} = \frac{\exp(-6.01047 + 0.34127 * 20)}{1+\exp(-6.01047 + 0.34127 * 20)} = 0.6931578$$



Se pueden ocupar las funciones *plogis* {stats} o *ilogit* {faraway} para determinar esta probabilidad

```
>ilogit( modeloLogit$coef[1]+ modeloLogit$coef[2]*20)
(Intercept) 0.6931578
```

Modelo probit

```
> modeloProbit<-glm(cbind(afectados,vivos)~concentracion,binomial(link="probit"), data=insecto1)
> summary(modeloProbit)

Call:
glm(formula = cbind(afectados, vivos) ~ concentracion, family = binomial(link = "probit"),
    data = insecto1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9997  -1.6451   0.1878   1.8272   2.3810

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.57902    0.42991  -8.325   <2e-16 ***
concentracion  0.20265    0.02254   8.989   <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: 138.001  on 9  degrees of freedom
Residual deviance:  37.552  on 8  degrees of freedom
AIC: 69.126

Number of Fisher Scoring iterations: 5
```

Aquí , al resolver la ecuación, lo que se obtiene un z-score

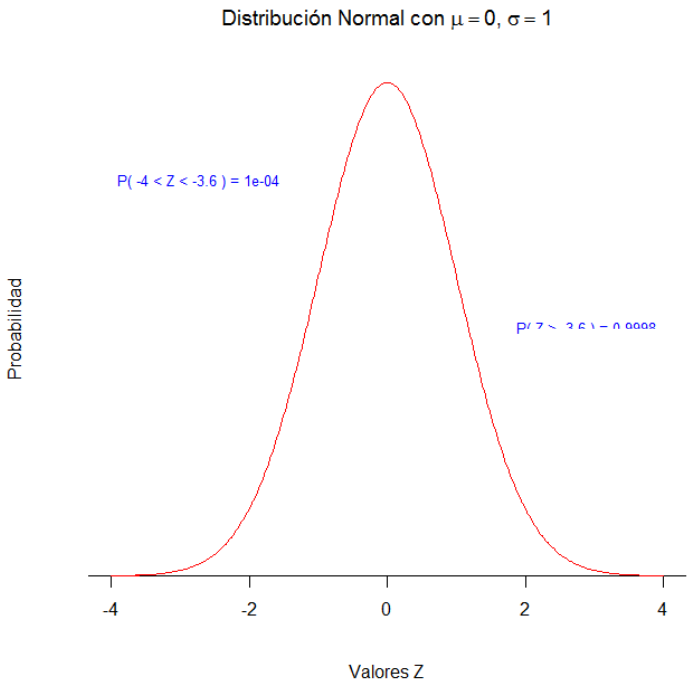
z- score= -3.57902 + 0.20265 *concentración

Interpretación de β_0 Cuando no hay insecticida en el ensayo (concentración de 0), el ϕ^{-1} (p) es -3.57

Interpretación de β_1 Por cada unidad de incremento en la concentración de insecticida, el ϕ^{-1} (p) incrementa en 0.20 unidades

ϕ^{-1} (p) = -3.57902 + 0.20265 *concentración

ϕ^{-1} (p) = -3.57902 + ~~0.20265~~ *0



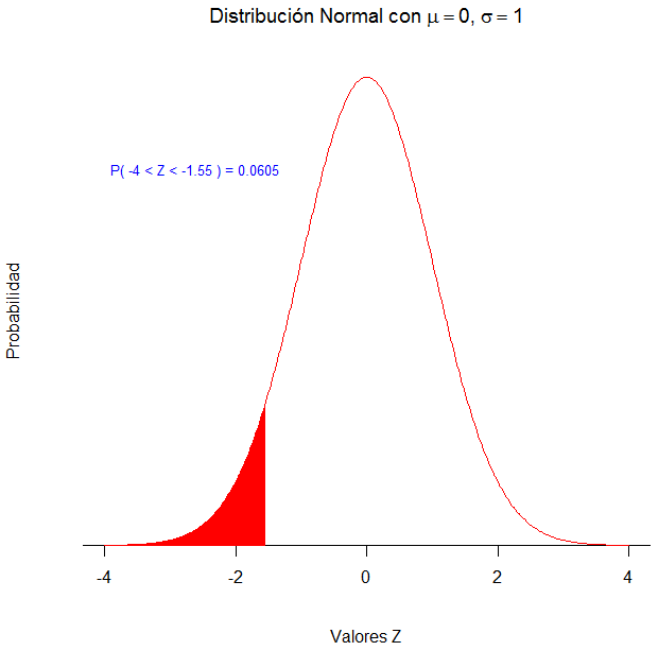

```
> modeloProbit$coef[1] + modeloProbit$coef[2]*0
(Intercept)
  -3.579018
> pnorm(modeloProbit$coef[1] + modeloProbit$coef[2]*0)
 (Intercept)
0.0001724441
```

Con una dosis de 0, el ϕ^{-1} (p) calculado es -3.57, y el área acumulada en la distribución normal desde -∞ hasta ese valor es 0.00017.

Por lo tanto, la probabilidad de morir a una dosis de 0 es ~ 0

ϕ^{-1} = Inverso de la distribución normal

ϕ^{-1} (p) es un valor Z o z score

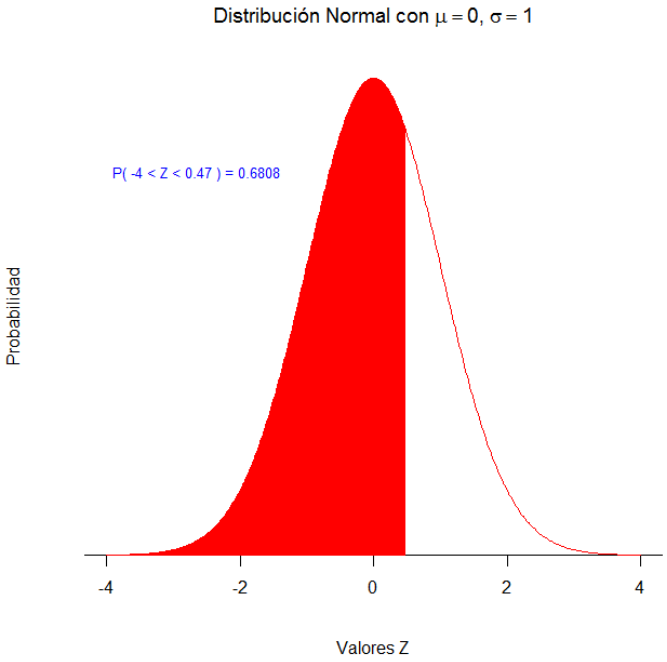



```
> modeloProbit$coef[1] + modeloProbit$coef[2]*10
(Intercept)
  -1.552526
```

```
> pnorm(modeloProbit$coef[1] + modeloProbit$coef[2]*10)
(Intercept)
 0.06026822
```

Con una dosis de 10, el ϕ^{-1} (p) calculado es -1.55, y el área acumulada desde -∞ hasta ese valor en la distribución Normal es 0.06.

Por lo tanto, usando liga probit, la probabilidad de morir a una dosis de 10 es 0.06

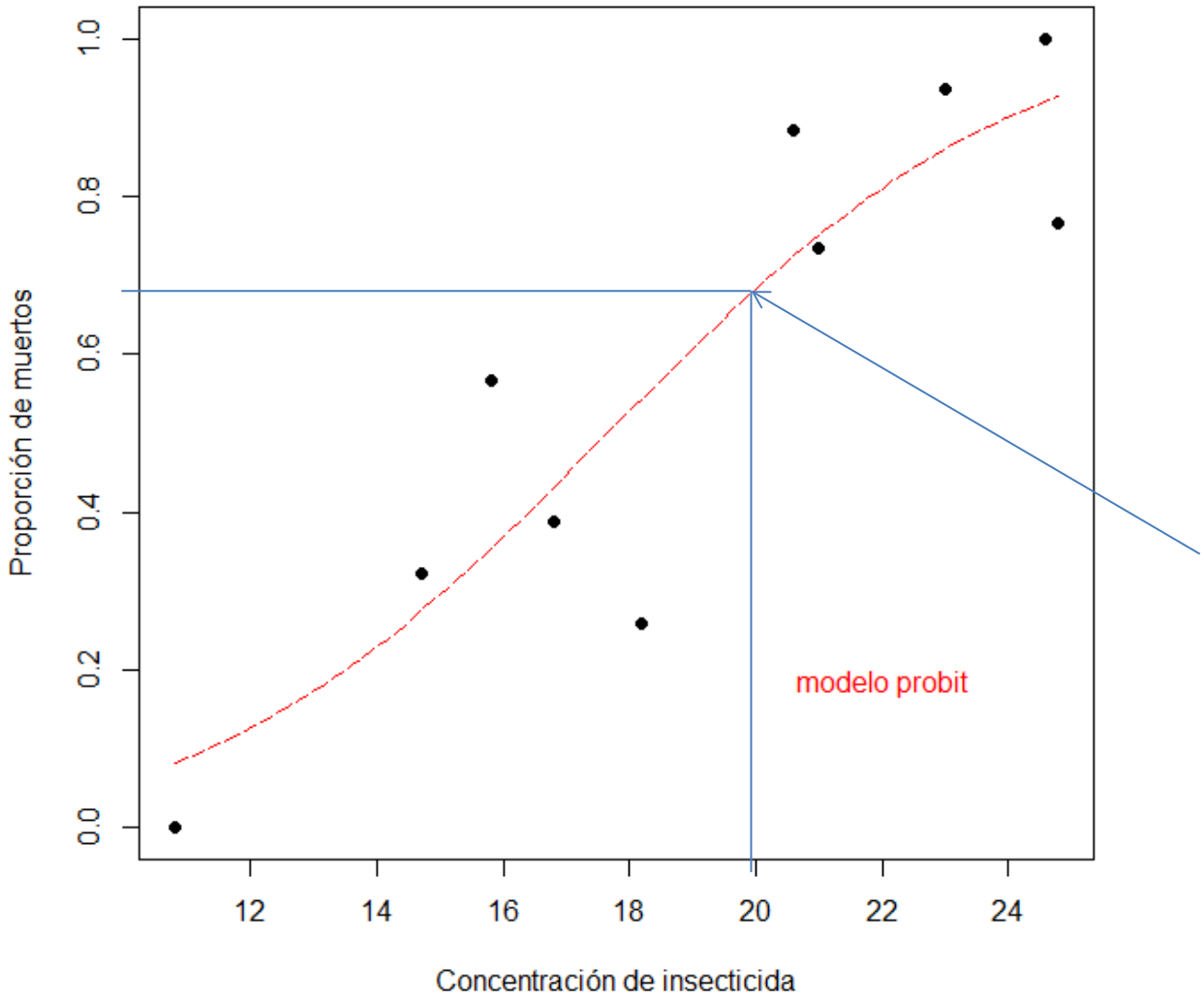



```
> modeloProbit$coef[1] + modeloProbit$coef[2]*20
(Intercept)
  0.4739657
```

```
> pnorm(modeloProbit$coef[1] + modeloProbit$coef[2]*20)
(Intercept)
  0.6822378
```

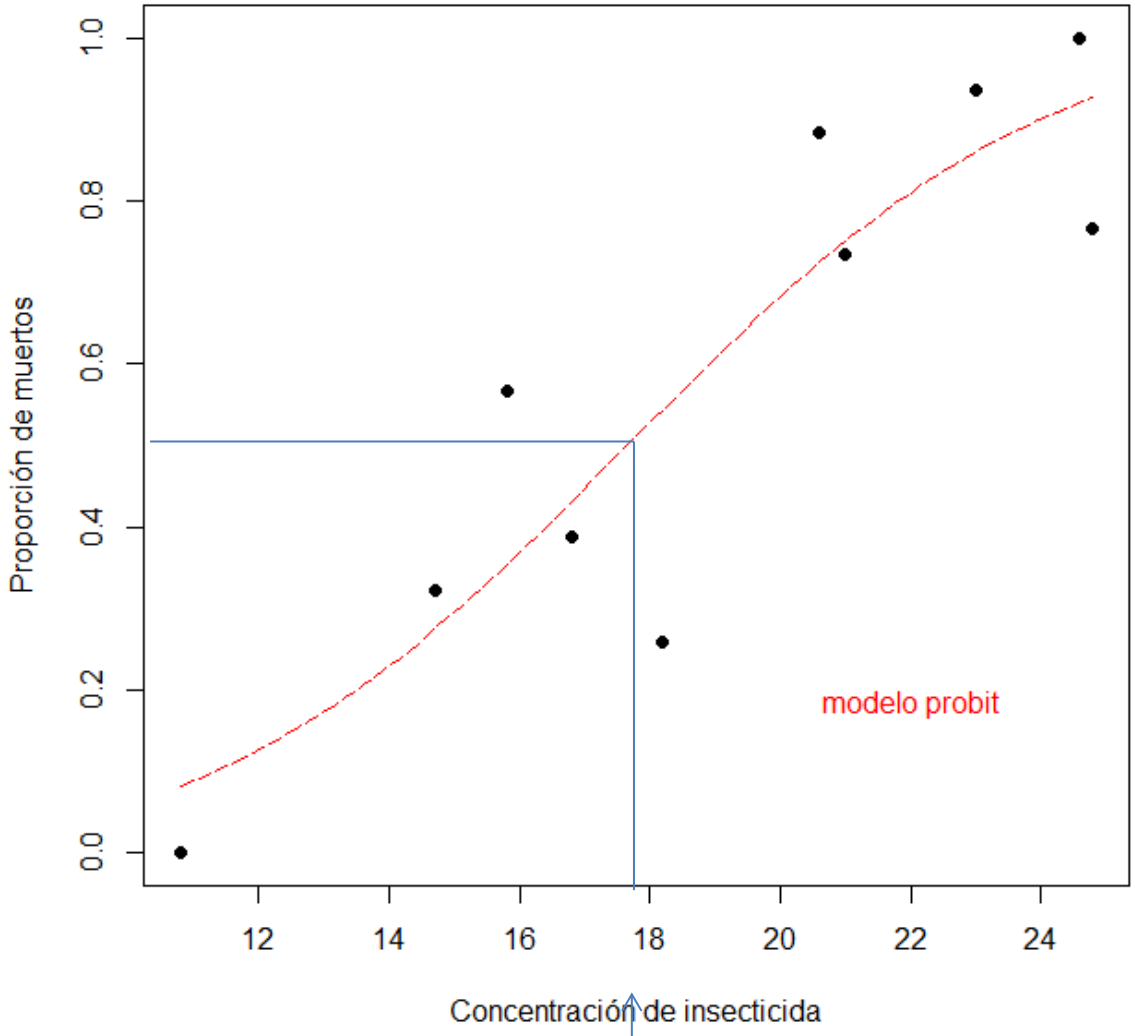
Con una dosis de 20, el ϕ^{-1} (p) calculado es 0.47, y el área acumulada desde -∞ hasta ese valor en la distribución Normal es 0.682

Por lo tanto, usando liga probit, la probabilidad de morir a una dosis de 20 es 0.682



Aquí podemos ver como va incrementando la probabilidad de morir conforme aumenta la concentración. A una dosis de 20, la probabilidad de morir es 0.682 (la que se había calculado en la diapositiva anterior)

```
> plot(insecto1$concentracion,prop_muertos, xlab="Concentración de insecticida", ylab="Proporción de muertos", pch
=16, col="black")
> curve(pnorm(modeloProbit$coef[1] + modeloProbit$coef[2]*x),add=T, lty=5,lwd=1.7, col="red")
> text(22,0.15,labels=c("modelo probit"), pos=3, col=c(2))
```



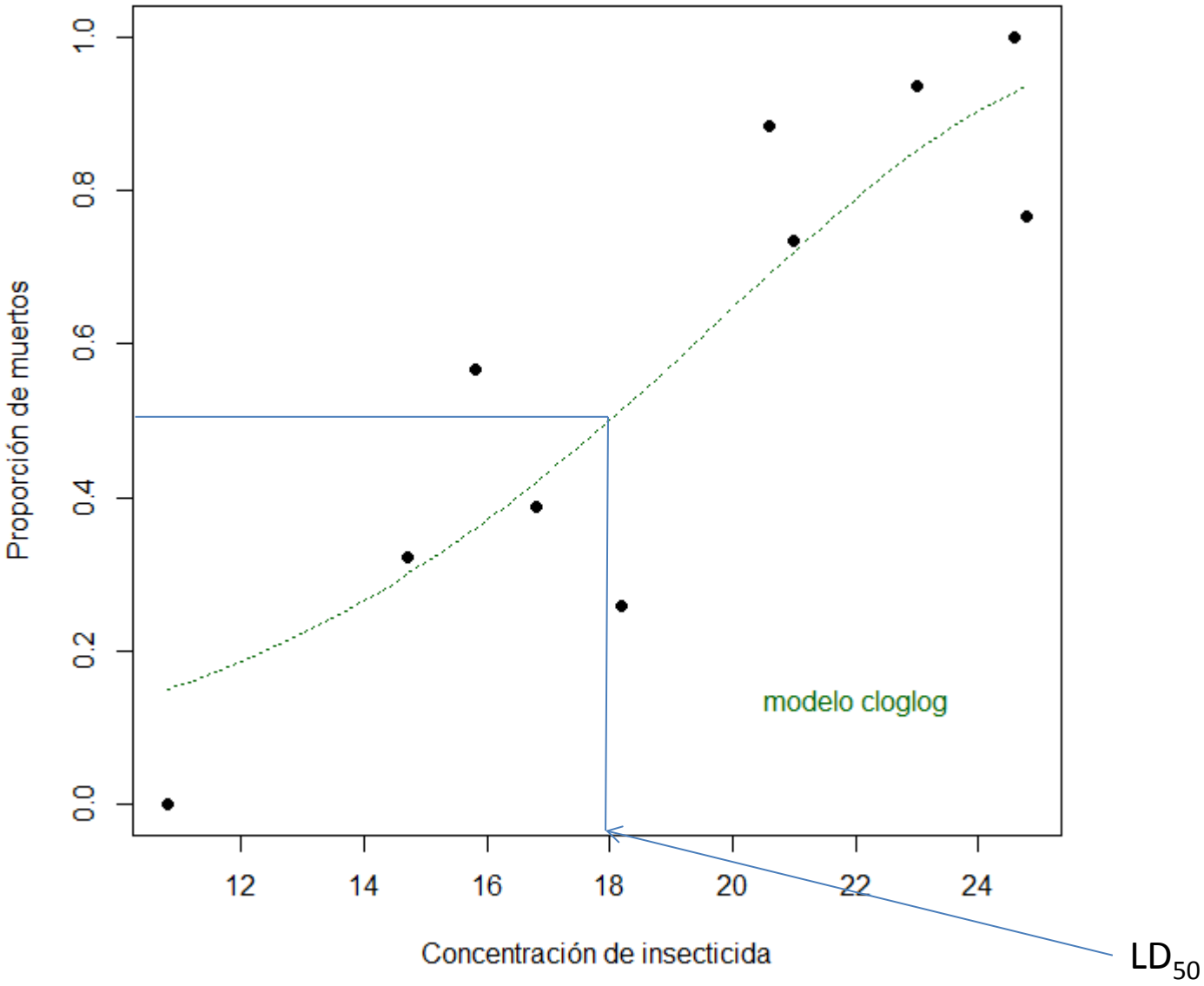

```
> (0-modeloProbit$coef[1])/modeloProbit$coef[2] #dosis LD50 con modelo probit
(Intercept)
   17.66115
```

Modelo cLoglog

```
> modeloCloglog<-glm(cbind(afectados,vivos)~concentracion,binomial(link="cloglog"), data=insecto1)
> summary(modeloCloglog)

Call:
glm(formula = cbind(afectados, vivos) ~ concentracion, family = binomial(link = "cloglog"),
    data = insecto1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9863  -2.1178   0.2222   1.9451   2.3360

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.00985    0.48644  -8.243   <2e-16 ***
concentracion  0.20244    0.02355   8.598   <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: 138.001  on 9  degrees of freedom
Residual deviance:  41.576  on 8  degrees of freedom
AIC: 73.149

Number of Fisher Scoring iterations: 5
```

Aquí , al resolver la ecuación, lo que se obtiene es log (-log(1-p)

cloglog(p) = log (-log(1-p) = -4.00985 + 0.20244 *concentración




```
> plot(insecto1$concentracion,prop_muertos, xlab="Concentración de insecticida", ylab="Proporción de muertos", pch=16,
col="black")
> curve(1-exp(-exp((modeloCloglog$coef[1]+ modeloCloglog$coef[2]*x))),add=T, lty=3,lwd=1.7, col="dark green")
> text(22,0.1,labels=c("modelo cloglog"), pos=3, col=c(81))
```

Interpretación de β_0 Cuando no hay insecticida en el ensayo (concentración de 0), el cloglog de la probabilidad de morirse (log (-log(1-p))es -4.00985

Interpretación de β_1 Por cada unidad de incremento en la concentración de insecticida, el cloglog de la probabilidad de morirse (log (-log(1-p)) incrementa 0.20244 unidades

```
#probabilidad de morir con una dosis de 20 en modelo cloglog
> 1-exp(-exp((modeloCloglog$coef[1]+ modeloCloglog$coef[2]*20)))
 (Intercept)
  0.6464782
```

```
(log(-log(0.5))-modeloCloglog$coef[1])/modeloCloglog$coef[2]
(Intercept)
   17.99673
```

Devianza

- Es una medida de desajuste en un Modelo Lineal Generalizado
- Medida de desajuste pues entre más grande es su valor, el ajuste del modelo es peor.
- Es un estadístico que compara la verosimilitud el modelo de interés con la verosimilitud el modelo saturado (el modelo en donde hay un parámetro para cada observación)

Devianza = $-2[L_M - L_S]$

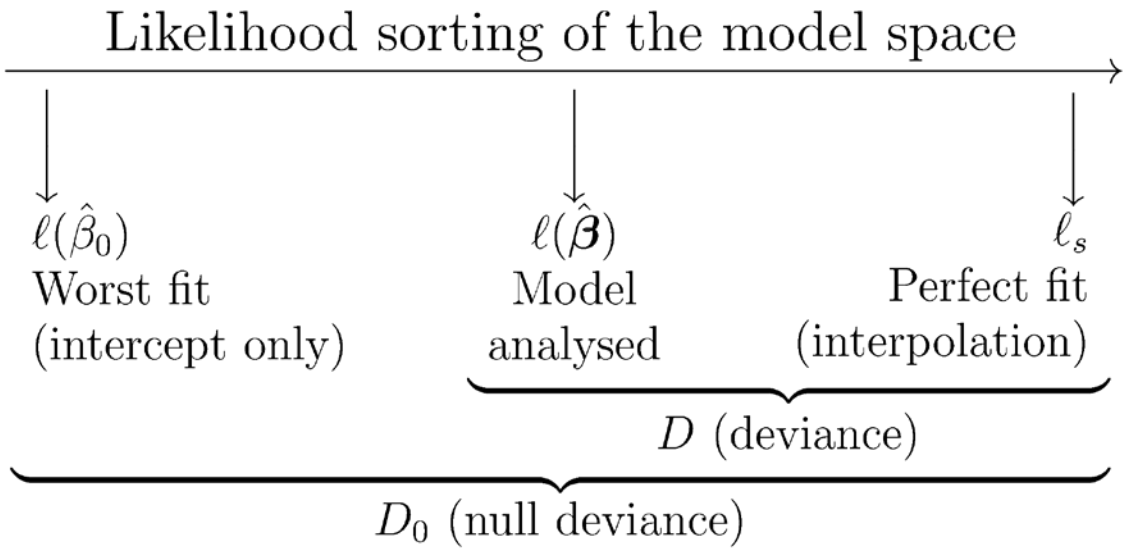
L_M log-verosimilitud maximizada de modelo de interés
L_S log-verosimilitud maximizada de modelo saturado

En R aparece al final del *summary* del modelo:

```
    Null deviance: 138.001  on 9  degrees of freedom
Residual deviance:  37.552  on 8  degrees of freedom
```

Indica que tan bien se predice la variable repuesta usando un modelo que incluye sólo β_0

Indica que tan bien se predice la variable repuesta usando un modelo que incluye también a las variables independientes



Efron's	$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{\pi}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$ $\hat{\pi}$ = model predicted probabilities	Efron's mirrors approaches 1 and 3 from the list above–the model residuals are squared, summed, and divided by the total variability in the dependent variable, and this R-squared is also equal to the squared correlation between the predicted values and actual values. When considering Efron's, remember that model residuals from a logistic regression are not comparable to those in OLS. The dependent variable in a logistic regression is not continuous and the predicted value (a probability) is. In OLS, the predicted values and the actual values are both continuous and on the same scale, so their differences are easily interpreted.

https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/

pseudoR2<-1 - (model$deviance/model$null.deviance)

Criterio de Información de Akaike (AIC)

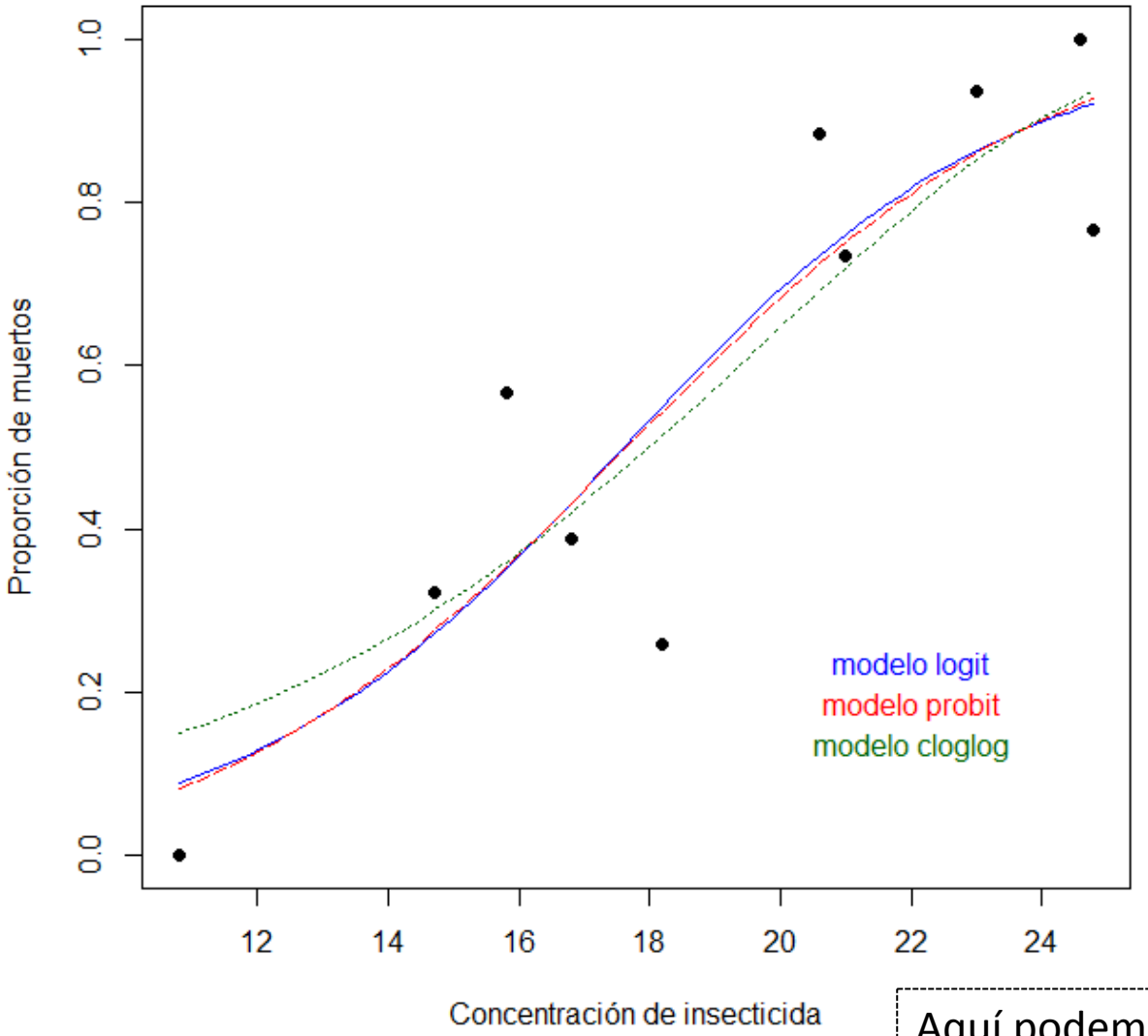
- Provee de un método para evaluar la calidad de modelos parecidos
- AIC = -2 (Verosimilitud del modelo) + 2 m

  m=número de parámetros
- Se basa en la devianza pero penaliza el hacer el modelo más complicado, es decir, al incluir un mayor número de parámetros, previene el incluir parámetros irrelevantes
- Se busca que el modelo explique muy bien a los datos, utilizando el menor número posible de parámetros.
- Al comparar varios modelos, conviene seleccionar el que tenga un menor valor.

```
    Null deviance: 138.001  on 9  degrees of freedom
Residual deviance:  37.552  on 8  degrees of freedom
AIC: 69.126
```

```
> AIC(modeloLogit) [1] 69.27027
```

En R aparece después de las devianzas.
O también, puede solicitarse con comando AIC




```
> AIC(modeloLogit,modeloProbit, modeloCloglog)
               df      AIC
modeloLogit     2 69.27027
modeloProbit    2 69.12561
modeloCloglog   2 73.14945
> AICtab(modeloLogit,modeloProbit, modeloCloglog)
               dAIC df
modeloProbit   0.0  2
modeloLogit    0.1  2
modeloCloglog 4.0  2
```

Aquí podemos apreciar, que aunque estos modelos no son tan diferentes, el modelo con menor AIC es el que tiene la liga probit. Hay 0.1 unidades de diferencia respecto al modelo Logit y 4 unidades de diferencia respecto al modelo cLoglog