

Ciclo de workshops de análisis de datos en epidemiología

10 de JUNIO 10 am

GLM para análisis de datos en fitopatología (I)

Dr. Juan Edwards Molina (INTA Balcarce)



Inscripción gratuita

<https://bit.ly/workshopanalisis>

10

15

20

25

30

35

Your X Variable

Proyectos INTA
1090 - 1074

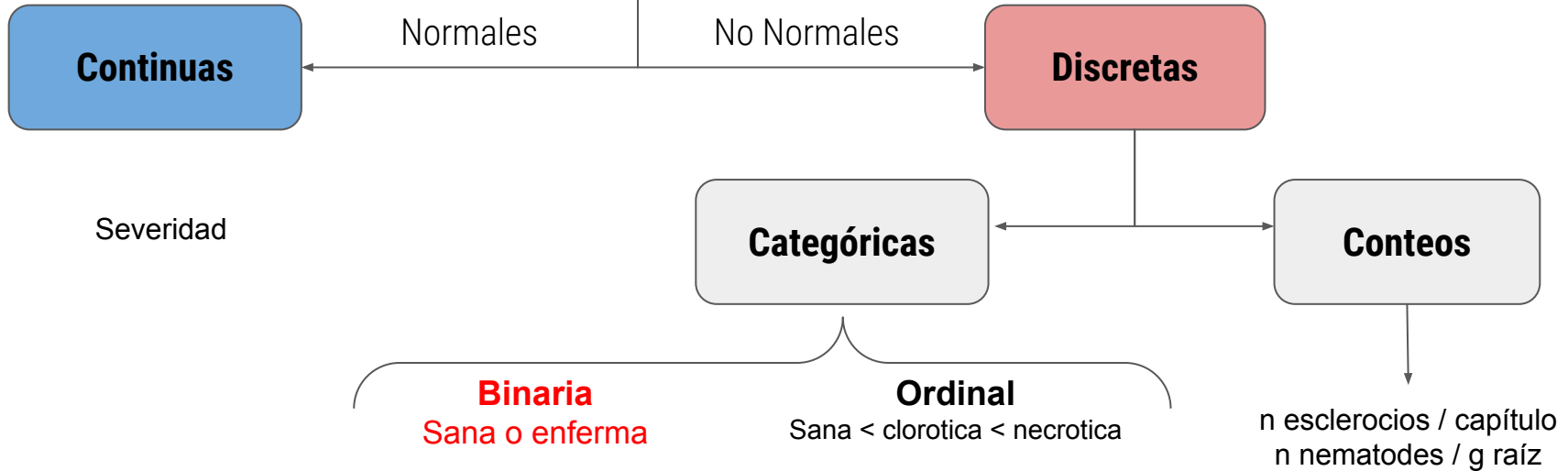
Programa Nacional
Protección Vegetal de INTA



Ministerio de Agricultura,
Ganadería y Pesca
Argentina

Evaluación visual de enfermedades

Tipo de variables



Normal

Binomial

Reg. ordinal o multinomial

Poisson

Modelos Lineales LM

Modelos lineales generalizados - GLM

¿Qué hacemos?

¿Adecuamos **nuestros datos** a las **técnicas analíticas**?
o mejor
¿las **técnicas analíticas** a **nuestros datos**?

Modelos lineales (LM)

Técnica estadística para modelar **relaciones lineales** entre una variable dependiente y una o múltiple variables independientes.

Supuestos:

- independencia de las observaciones
- homocedasticidad de la varianza
- normalidad de los residuos

Regresión lineal

Homogeneity of Variance
Reference line should be flat and horizontal

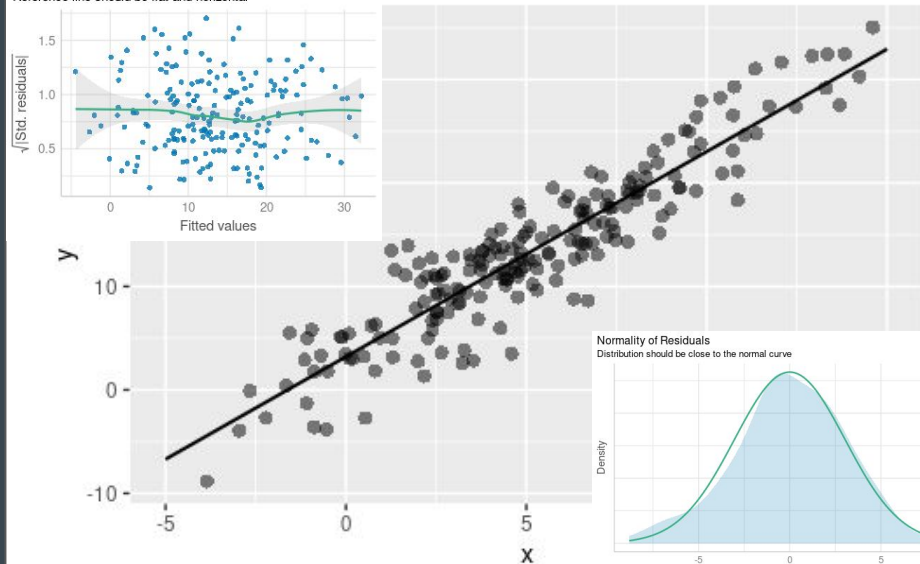


Diagram illustrating the components of the linear regression equation:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Labels and components:

- Variable dependiente** points to Y_i .
- Intercepto** points to β_0 .
- Pendiente** points to β_1 .
- Variable independiente** points to X_i .
- Error** points to ε_i .
- componentes lineales (SISTEMÁTICO)** is indicated by a bracket under $\beta_0 + \beta_1 X_i$.
- componente aleatorio** is indicated by a bracket under ε_i .

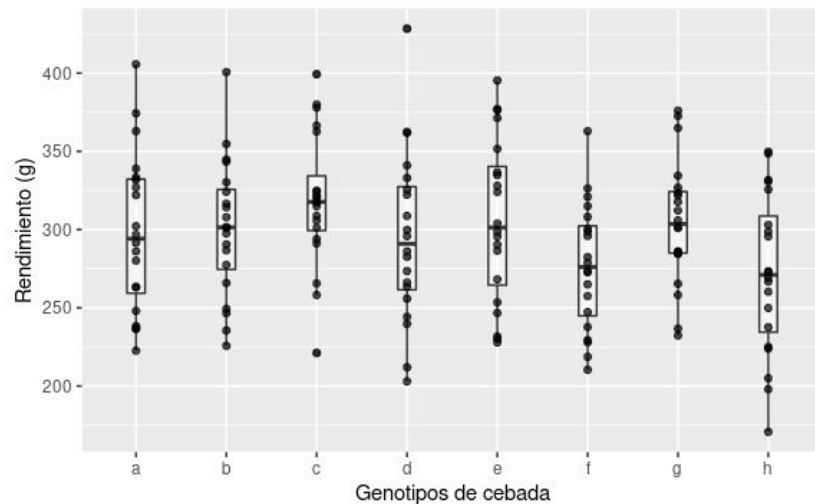
Modelos lineales (LM)

Técnica estadística para modelar **relaciones lineales** entre una variable dependiente y una o múltiple variables independientes.

Supuestos:

- independencia de las observaciones
- homocedasticidad de la varianza
- normalidad de los residuos

DBCA



Variable respuesta
para el i-trt en el j-bk

Constante
Media gral.

Efecto del
i-trt

Efecto del
j-bk

Residual

$$Y_{ij} = \theta + \tau_i + b_j + e_{ij}$$

$$b_j \sim N(0, \sigma_b^2) \quad e_{ij} \sim N(0, \sigma_e^2)$$

Modelos lineales generalizados (GLM)

Generalización flexible de los LM que admite variables respuesta con distribución de error distinta de una normal, al permitir que el componente lineal se relacione con la variable respuesta a través de una **función de enlace** (link).

Una propiedad de distribuciones no-normales, en general, es que la **varianza** de la distribución es **función de la media**. Esto significa que los niveles de un factor (tratamientos) tendrán diferentes varianzas (violación a los supuestos de los LM : varianzas constantes)

Supuestos

- Independencia de Y (como fueron tomados los datos? qué tipo?)
- Correcta función de enlace
- Ausencia de observaciones influyentes

LM

Componente sistemático (pred. lineal)

$$\mu = \beta_0 + \beta_1 x$$

Componente aleatorio

$$y_i = \text{Normal}(\mu_i)$$

GLM

Componente sistemático (pred. lineal)

$$\eta = \beta_0 + \beta_1 x$$

Función de enlace

$$\eta = \text{link}(\mu)$$

Componente aleatorio

$$y_i = \text{distribución}(\mu_i)$$

GLM para variable binomial

Regresión logística

Componente sistemático

$$\eta = \beta_0 + \beta_1 x$$

Función de enlace

$$\eta = \text{logit}(\mu_i) = \log(\mu_i / 1 - \mu_i)$$

Componente aleatorio

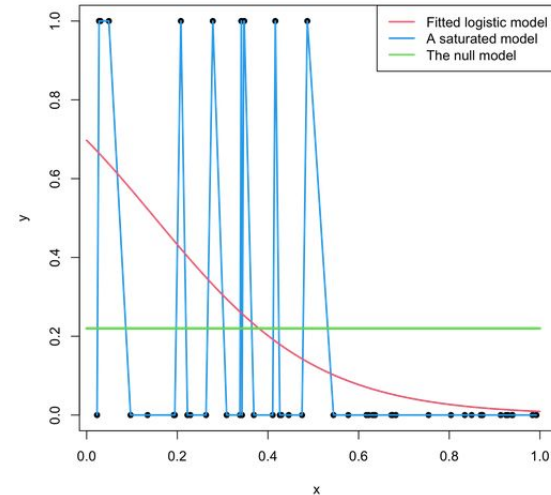
$$y_i = \text{binomial}(\mu_i)$$

GLM para variable binomial

Análisis de deviance

Generalización del análisis de la varianza para los GLM obtenido para una secuencia de modelos anidados (cada uno incluyendo más términos que los anteriores).

La deviance mide la desviación del GLM con respecto a un modelo perfecto para la muestra (modelo saturado), la cual se ajusta perfectamente a los datos



Modelo lineal generalizado para variable conteo

Regresión poisson

Componente sistematico

$$\eta = \beta_0 + \beta_1 x$$

Función de enlace

$$\eta = \log(\mu)$$

Componente aleatorio

$$y_i = \text{poisson}(\mu_i)$$

Variables Binomiales I

Incidencia

- Nivel intra-planta
 - frutas de naranja con antracnosis [10.1094/PDIS-01-19-0068-RE](https://doi.org/10.1094/PDIS-01-19-0068-RE)
 - virus: ToCV en hojas de tomate (elisa⁻=0 ; elisa⁺=1) [10.1094/phyto-06-18-0203-r](https://doi.org/10.1094/phyto-06-18-0203-r)
- Nivel parcela
 - Vainas de maní fuera del estándar comercial (No=1; Si=0) [10.1016/j.cropro.2020.105403](https://doi.org/10.1016/j.cropro.2020.105403)
- Nivel lote
 - CABMV virus en plantas de maracuya (0-1) [10.1111/ppa.13054](https://doi.org/10.1111/ppa.13054)

Prevalencia

- Ausencia / presencia de phomopsis del girasol en un lote

Otros

- ¿Se solventó el tratamiento fungicida? No=0; Si=1
- ¿Se alcanzó el umbral de aplicación? No=0; Si=1

Medidas de frecuencia

Medida de efecto

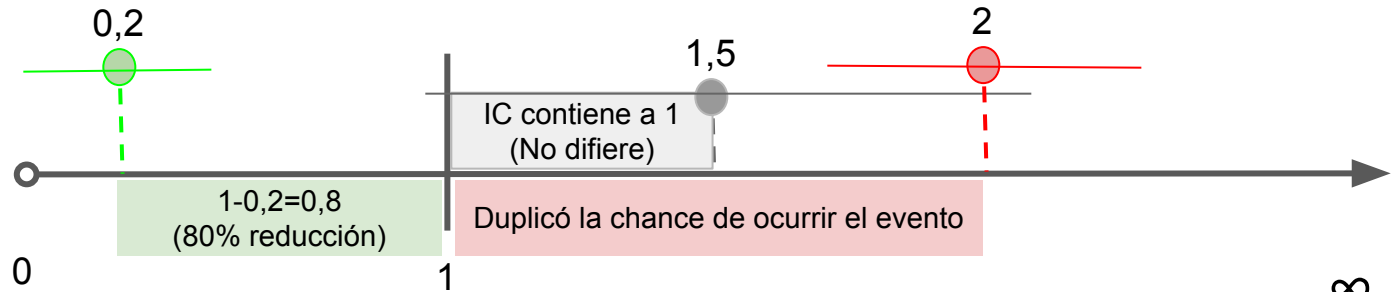
Probabilidad	$\frac{\text{Nº de eventos favorables}}{\text{Nº de eventos \textbf{totales}}}$
Riesgo	Probabilidad de que ocurra un evento "negativo"
Odds	$\frac{\text{Nº de eventos favorables}}{\text{Nº de eventos \textbf{desfavorables}}}$

[0;1]

Odds = $p / (1-p)$ = p de que ocurra / p de que no ocurra

Odds ratio = Odds_{trat} / Odds_{control}

Interpretación



log odds	$\text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \alpha + \beta_1 x_1 + \cdots + \beta_n x_n$	Así se estiman los coeficientes con GLM
odds	$\frac{p}{1 - p} = \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})$	así podemos reportar los efectos de las predictoras
p(Y X)	$\Pr(y_i = 1) = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})}$	predecimos la probabilidad para valores de X

R - Outline

1. Entendiendo la naturaleza binaria de la incidencia
2. DBCA (análisis de deviance) - *data phomopsis*
 - a. Ajuste mediante LM y GLM - comparación
 - b. Diagnósticos
 - c. Interpretacion de coeficientes (log OR, OR, p)
3. Regresión logística - *data maracuyá*
 - a. Single / multiple-point assessment
 - b. Curva de progreso de la incidencia
 - c. Predicciones

Conclusiones

1. Ajustamos la técnica de análisis a la naturaleza de nuestros datos
 - a. Vimos que no llegamos a conclusiones similares
2. Los modelos mixtos nos permiten lidiar con la violación de algunos supuestos de los GLM (independencia de las observaciones)
3. Actualmente hay paquetes de R para realizar el workflow completo de análisis (sin recurrir a cálculos manuales)

Sintaxis en R

Efectos fijos	Efectos mixtos
<ul style="list-style-type: none">• {stats} lm	<ul style="list-style-type: none">• {lme4} lmer• {nlme} lme +permite modelar varianza
<ul style="list-style-type: none">• {stats} glm +family=quasibinomial	<ul style="list-style-type: none">• {lme4} glmer• {glmmTMB} glmmTMB +tienen muchas alternativas de distribuciones• {MASS} glmmPQL (Penalized Quasi-Likelihood)

Distribución Binomial - propiedades

Y: Número de individuos con cierta carácter (EXITOS, ej., enfermedad) en una unidad experimental o muestral (ej., parcela, planta) – respuesta

n: Número de individuos observados para el carácter (ej., plantas)

p: Parámetro de localización: probabilidad de un carácter, como una enfermedad (ej., probabilidad de que una hoja, planta, etc., está enferma) (análogo a μ de normal)

- Para una simple muestra aleatoria de n plantas, la incidencia de la enfermedad (como proporción) es una estimación de **p**
- La varianza de la distribución condicional de Y es **$np(1-p)$** , completamente definida por **n** y **p**

Cuanto mayor n , $\text{Bin}(p, n)$ se aproxima a la distribución normal para una muestra simple, con media **np** , y varianza **$np(1-p)$**

Recursos

[Workshop 6: Generalized linear models](#)

[Chapter 5 Generalized linear models | Notes for Predictive Modeling](#)

[Logistic regression](#)

[Regresion Logistica: Interpretacion de Coeficientes. Pronosticos.](#)

<https://stats.oarc.ucla.edu/r/dae/logit-regression/>

[Using R to make sense of the generalised linear model | BARELY SIGNIFICANT](#)

https://rpubs.com/benhorvath/logistic_regression

https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704-ep713_multivariablemethods/BS704-EP713_MultivariableMethods4.html

http://umh1480.edu.umh.es/wp-content/uploads/sites/44/2013/02/tema_5_1.pdf

<http://glmm.wikidot.com/examples>

<https://stats.stackexchange.com/questions/185491/diagnostics-for-generalized-linear-mixed-models-specifically-residuals>