Modelos Bionomiales

Modelos Estadísticos Avanzados

Santiago Benitez-Vieyra

Distribución Binomial.

1. Modelos para datos de presencia-ausencia.

Proceso de Bernoulli

serie de “experiencias” independientes donde la respuesta puede ser sólo de dos categorías.

- Es una respuesta binaria cualquier variable que pueda ser clasificada en sólo dos niveles: presencia/ausencia, categoría1/categoría2, morfo1/morfo2, … en general éxito/fracaso.
- La media de una distribución binaria es igual a la probabilidad de ocurrencia de éxitos.
- La varianza de una distribución binaria es igual a la probabilidad de ocurrencia de éxitos por la probabilidad de ocurrencia de fracasos.

Proceso de Bernoulli

serie de experiencias independientes donde la respuesta puede ser sólo de dos categorías.

$$P(y = 1) = p$$

$$P(y = 0) = 1 - p$$

$$E(y) = p$$

Datos binomiales agrupados

- Cada respuesta está constituida por más de un proceso de Bernoulli.
- Aplican a datos de PROPORCIONES, en general conocemos el número de éxitos y el número total de procesos de Bernoulli por sujeto (fracasos = total – éxito).

$$E(y) = Np$$

La variable respuesta se distribuye como Binomial

$$Y_i \sim B(N, p)$$

El enlace canónico es logit

$$\eta = log(\frac{\mu_i}{1 - \mu i})$$

Por lo que el valor esperado es igual al logit inverso.

$$\mu_i = \frac{e^{\eta}}{1 + e^{\eta}}$$

Colocando las variables predictoras en la parte sistemática…

$$\mu_i = \frac{e^{\beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}}}{1 + e^{\beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}}}$$

Chances (odds)

$$logit(p(y)) = log\frac{p(y)}{1 - p(y)}$$

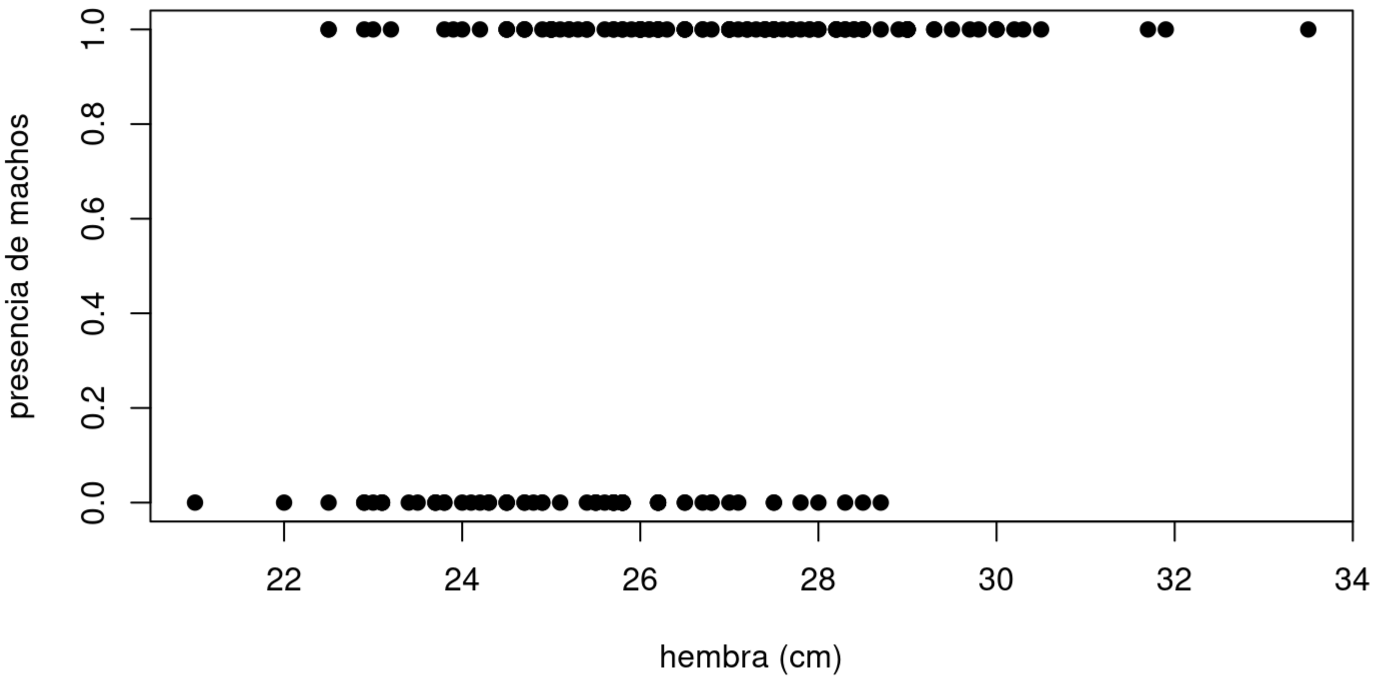
Donde $p(x)$ es la probabilidad de éxito y $1 - p(x)$ es la probabilidad de fracaso.

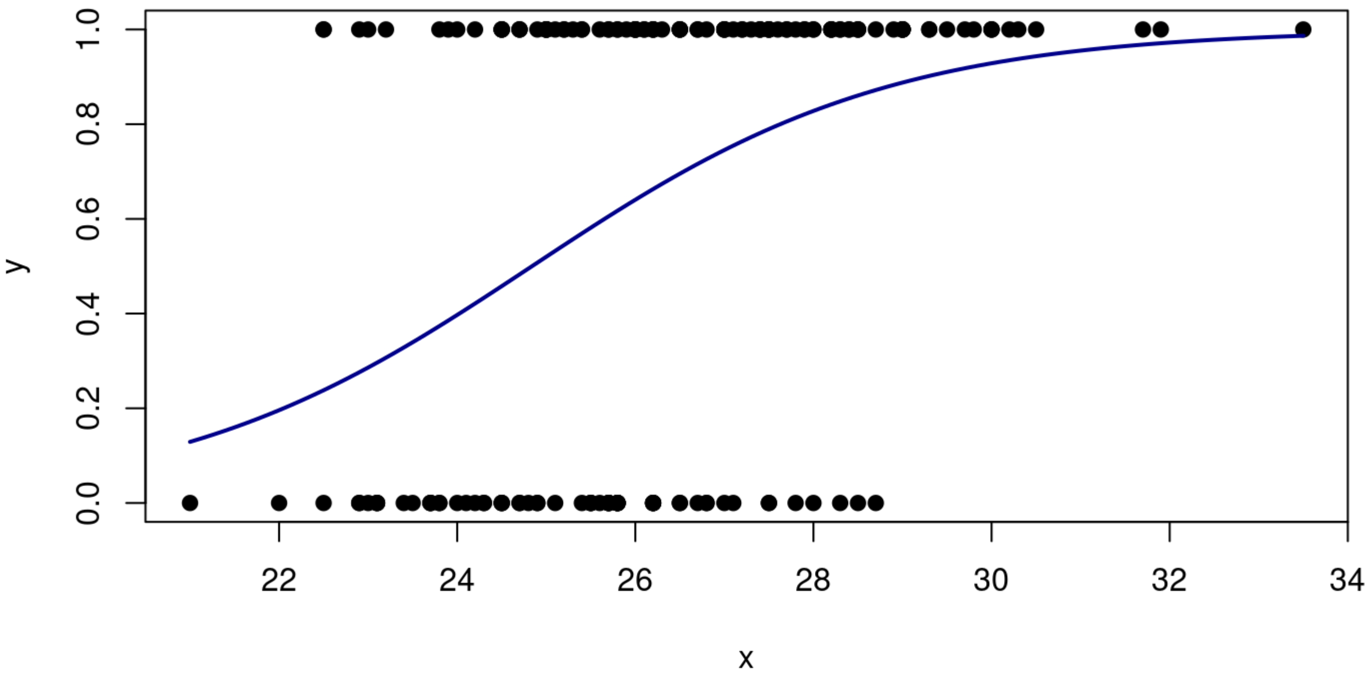
Chances (odds)

$$log \frac{p(y)}{1-p(y)} = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}$$

La respuesta es lineal en el espacio logit.

El cangrejo herradura


presencia de machos
0.0
0.2
0.4
0.6
0.8
1.0
22
24
26
28
30
32
34
hembra (cm)


0.0
0.2
0.4
0.6
0.8
1.0
y
22
24
26
28
30
32
34
x

Análisis de la Devianza

```
fit <- glm(y ~ x, data = dat, family = binomial)
anova(fit, test = "Chisq")
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: y

Terms added sequentially (first to last)

     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                   172     225.76
x     1   31.306       171     194.45 2.204e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test de cociente de verosimilitudes

```
fit0 <- glm(y ~ 1, data = dat, family = binomial)
fit1 <- glm(y ~ x, data = dat, family = binomial)
anova(fit0, fit1, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: y ~ 1
Model 2: y ~ x
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       172     225.76
2       171     194.45  1   31.306 2.204e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretación de parámetros

```
exp(fit$coefficients)
```

```
 (Intercept)            x
4.326214e-06 1.644162e+00
```

- La chance de tener éxito se incrementa 64.41% por cada unidad de aumento en la variable x.
- La probabilidad de tener éxito respecto a la probabilidad de no tenerlo se incrementa un 64.41%.
- El cociente de chances (odds ratio) es de 1.64 entre dos observaciones que difieran en una unidad.

La dosis letal 50%

- La dosis “letal” 50 toma su nombre de los ensayos de dosis-respuesta, en los cuales se aplica una dosis creciente de determinada droga y se examina una respuesta (que habitualmente es muerte/supervivencia).
- La LD50 indica el valor de la variable independiente a la cual obtenemos el mismo número de éxitos y de fracasos.

$$LD_{50} = -\beta_0/\beta_1$$

Matriz de confusión

```r
pred <- predict(fit, type = "response") #training or test?
table(data = as.numeric(pred >= 0.5), reference = dat$y) #umbral
```

```
    reference
data  0  1
   0 27 16
   1 35 95
```

ver también `confusionMatrix` del paquete `caret`

T. Fawcett / Pattern Recognition Letters 27 (2006) 861–874

		True class: p	True class: n
Hypothesized class	Y	True Positives	False Positives
	N	False Negatives	True Negatives
Column totals:		P	N

$$\text{fp rate} = \frac{FP}{N} \qquad \text{tp rate} = \frac{TP}{P}$$

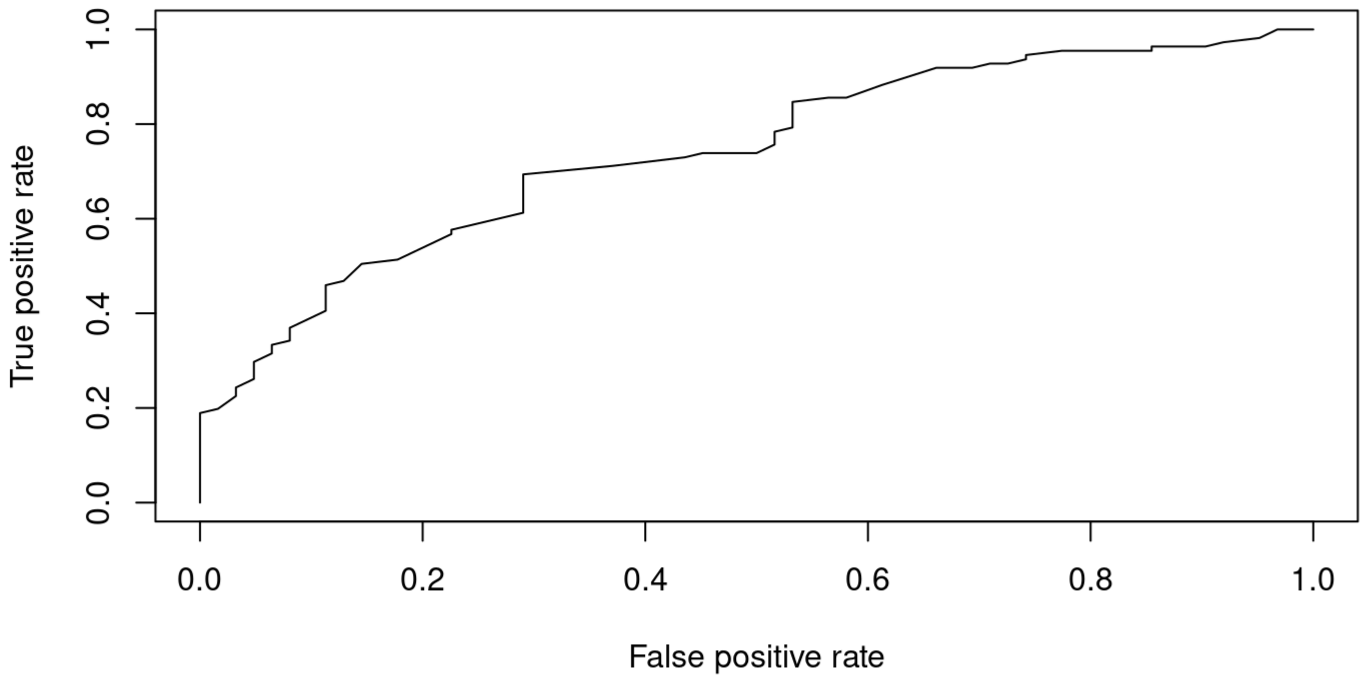
$$\text{precision} = \frac{TP}{TP+FP} \qquad \text{recall} = \frac{TP}{P}$$

$$\text{accuracy} = \frac{TP+TN}{P+N}$$

$$\text{F-measure} = \frac{2}{1/\text{precision}+1/\text{recall}}$$

Fig. 1. Confusion matrix and common performance metrics calculated from it.

ROC y AUC



cada punto corresponde a una matriz de confusión con diferente umbral.

VER APP

Sobredispersión

- La sobredispersión es provocada por variación al azar **NO EXPLICADA** en la variable respuesta.
- Habitualmente se observa una importante devianza residual, que permanece incluso si se incorporan variables al modelo.
- Esta variabilidad provoca que la relación media-varianza esperada para la familia no se cumpla.
- Debe distinguirse de la sobredispersión aparente , debida a variables o interacciones faltantes, efectos no lineales no considerados, outliers en la variable respuesta o errores en la elección del enlace.

Uno de los supuestos de un MLG es que la variable dependiente se ajusta a una de las funciones de la familia exponencial. Cada una de estas familias está caracterizada por una relación media-varianza específica.

Distribución	Posición	Dispersión
Poisson	$E(Y) = \mu$	$var(Y) = \mu$
Binomial	$E(Y) = N\pi$	$var(Y) = N\pi(1 - \pi)$

Sin embargo, a veces la variabilidad observada es mayor en una proporción ϕ respecto a lo esperado. Si $\phi > 1$ entonces existe sobredispersión

Distribución	Posición	Dispersión
Poisson	$E(Y) = \mu$	$var(Y) = \phi\mu$
Binomial	$E(Y) = N\pi$	$var(Y) = \phi N\pi(1-\pi)$

Veremos más sobre sobredispersión en la clase sobre modelos de conteos. Por lo pronto utilizaremos la estimación de $\phi > 2$ como límite para modelos sobredispersos.

Sólo los modelos binomiales agregados pueden tener sobredispersión. Ni los binarios ni modelos donde el modelo saturado esté en consideración pueden tenerla.

END