

Respuesta conteos

Guillermo Ayala Gallego

Respuesta conteos

Guillermo Ayala Gallego

2025-02-02

Regresión Poisson

La distribución Poisson

- La función de probabilidad de una distribución de Poisson es $f(y; \mu) = \frac{e^{-\mu} \mu^y}{y!}$, con $y = 0, 1, \dots$.
- La podemos expresar como un elemento de la familia exponencial natural $f(y; \mu) = \frac{e^{-\mu} \mu^y}{y!} = \exp\{y \ln \mu - \mu - \ln(y!)\}$, tomando $(\theta_i = \ln \mu_i)$, $(b(\theta_i) = \exp\{\theta_i\})$, $(a(\phi) = 1)$, $(c(y_i, \phi) = -\ln(y_i!))$.

Ecuaciones de verosimilitud

- Son las siguientes $\sum_{i=1}^n (y_i - \mu_i) x_{ij} = 0$.
- La desviación (escalada y sin escalar) y con término constante es $D(\mathbf{y}; \hat{\mu}) = 2 \sum_{i=1}^n y_i \ln \frac{y_i}{\hat{\mu}_i}$.
- Adopta la expresión $D(\mathbf{y}; \hat{\mu}) = 2 \sum \text{Observado} \ln \left(\frac{\text{Observado}}{\text{Esperado}} \right)$.

tamidata2:PRJNA218851

Datos

```
1 pacman::p_load(SummarizedExperiment)
2 data(PRJNA218851, package="tamidata2")
1 table(colData(PRJNA218851)[, "Stage"])
```

Cancer	Metastasis	Normal
18	18	18

- Nos fijamos en el gen que ocupa la fila 1000.
- Construimos un `data.frame` en donde incluimos la variable `Stage` y el conteo correspondiente a este gen.

```
1 df = data.frame(count = assay(PRJNA218851)[1000,],
2                 Stage=colData(PRJNA218851)[,"Stage"])
3 head(df)
```

	count	Stage
SRR975551Aligned.out.sam.bam	539	Cancer
SRR975552Aligned.out.sam.bam	563	Cancer
SRR975553Aligned.out.sam.bam	1018	Cancer
SRR975554Aligned.out.sam.bam	393	Cancer
SRR975555Aligned.out.sam.bam	398	Cancer
SRR975556Aligned.out.sam.bam	672	Cancer

Regresión Poisson

- Ajustamos un modelo loglineal de Poisson.

```
1 fit = glm(count ~ Stage, family = poisson(link = log), data = df)
2 summary(fit)
```

Call:

```
glm(formula = count ~ Stage, family = poisson(link = log), data = df)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	6.729957	0.008146	826.14	<2e-16 ***
StageMetastasis	-0.306800	0.012512	-24.52	<2e-16 ***
StageNormal	0.429249	0.010467	41.01	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 13236.9 on 53 degrees of freedom
Residual deviance: 8723.7 on 51 degrees of freedom
AIC: 9189.2

Number of Fisher Scoring iterations: 4

- No es adecuado por varias razones:
 - Distintas profundidades de secuenciación.
 - Posiblemente sobre dispersión.

- No lo podemos usar.
- Vemos la diferencia de las desviaciones.

```
1 fit$null.deviance - fit$deviance
[1] 4513.206
```

Sobre dispersión

- Vamos a estimar la sobre dispersión.

```
1 fit1 = glm(count ~ Stage, family = quasipoisson(link = log), data = df)
2 summary(fit1)$dispersion
[1] 198.0956
```

Regresión binomial negativa

Distribución binomial negativa

- La función de probabilidad es $f(y; \phi, \mu) = \frac{\Gamma(y+\phi)}{\Gamma(\phi)\Gamma(y+1)} \left(\frac{\phi}{\mu+\phi}\right)^\phi \left(1 - \frac{\phi}{\mu+\phi}\right)^y$ con $(y = 0, 1, 2, \dots)$ donde (ϕ) y (μ) son los parámetros.
- Se tiene que $E(Y) = \mu$, $\text{var}(Y) = \mu + \frac{\mu^2}{\phi}$.
- El parámetro $(1/\phi)$ es un **parámetro de dispersión**.

Ejemplo

```
1 library(MASS)
2 fit = MASS::glm.nb(count~Stage, data=df)
3 summary(fit)
```

Call:

```
MASS::glm.nb(formula = count ~ Stage, data = df, init.theta = 5.191786539,
  link = log)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	6.7300	0.1038	64.858	< 2e-16 ***
StageMetastasis	-0.3068	0.1468	-2.090	0.03666 *
StageNormal	0.4292	0.1467	2.927	0.00343 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(5.1918) family taken to be 1)

Null deviance: 81.344 on 53 degrees of freedom
Residual deviance: 55.733 on 51 degrees of freedom
AIC: 796.61

Number of Fisher Scoring iterations: 1

Theta: 5.192
Std. Err.: 0.976

2 x log-likelihood: -788.611