Modelo Lineal Generalizado (GLM)

Técnica de Modelado Paramétrico

¿ Para que sirve GLM ?

- Para Predecir la variable respuesta (Y) en contextos de NO Normalidad y Heterocedasticidad.
- Para Inferir el efecto de covariables (X) en la respuesta (Y).
- Para Modelar la DISTRIBUCION de Y condicional a X.

El Concepto de GLM !!!

- La idea principal (! muy poderosa !) es la de tomar un modelo paramétrico probabilístico, y “modelar” los parametros del mismo en función de ciertos “features” obsrvables.

Caso particular: Modelo Lineal

$$Y \sim N(\mu(x_1, x_2, x_3, \ldots, x_n), \sigma^2)$$

$$\mu(x_1, x_2, x_3, \ldots, x_n) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n$$

Caso general

Variable aleatoria

Vector de Features

$$Y \sim F(\theta_\beta(\mathbf{x}))$$

Parámetros a ser estimados

Modelo probabilístico

Regresión Logística

Técnica de Clasificación

La Idea

Variable
Aleatoria

$$y \sim binomial(1, p) \quad \Longrightarrow \quad y \begin{cases} 1 & p \\ 0 & 1 - p \end{cases}$$

La probabilidad
(el parámetro) es
constante

Permitir que el parámetro
dependa de Features (**x**) que
caracterizan a la observación (i)

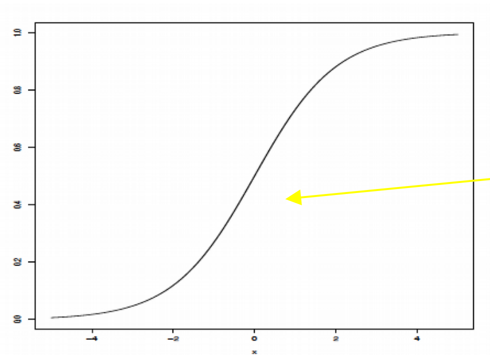
$$\boxed{p \rightarrow p(\mathbf{x}_i)}$$

El Modelo

Evento dicotómico

Vector de Features

$$P(Y=1|X) = \text{expit}\left(\beta_0 + \beta_1 X_1 + \ldots \beta_p X_p\right)$$



$$\text{expit}(t) = \frac{e^t}{1+e^t}$$

Expresión Lineal

$$\text{logit}(P(Y=1|X)) = \beta_0 + \beta_1 X_1 + \ldots \beta_p X_p$$

Función “Link” que mapea el [0,1] a los Reales

$$\text{logit}(t) = \log\left(\frac{t}{1-t}\right)$$

El Modelo en función de las "Odds"

Vector de Features

$$\frac{P(Y=1|X)}{1-P(Y=1|X)} = e^{\beta_0+\beta_1 X_1+\ldots\beta_p X_p}$$

Odd o Chance del evento

Componente lineal

La Estimación (Fisher Scoring) Basada en el Método de Máxima Verosimilitud

Verosimilitud

Parámetros a ser estimados

$$L(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i},$$

Log-Verosimilitud

$$\log L(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)$$

Probabilidad del evento, dependiente de los Features

$$\pi_i = \text{expit}\left(\beta_0 + \beta_1 X_1 + \ldots \beta_p X_p\right).$$

Extensión a K clases

Regresión Logística Multinomial

(Tomando 1 clase como referencia)

$$\ln \frac{\Pr(Y_i = 1)}{\Pr(Y_i = K)} = \boldsymbol{\beta}_1 \cdot \mathbf{X}_i$$

$$\ln \frac{\Pr(Y_i = 2)}{\Pr(Y_i = K)} = \boldsymbol{\beta}_2 \cdot \mathbf{X}_i$$

$$\cdots\cdots$$

$$\ln \frac{\Pr(Y_i = K-1)}{\Pr(Y_i = K)} = \boldsymbol{\beta}_{K-1} \cdot \mathbf{X}_i$$

K-1 conjuntos de coeficientes

Puedo despejar y calcular las probabilidades por clase !

Clase de referencia

Como quedan las probabilidades ?

$$\Pr(Y_i = 1) = \frac{e^{\boldsymbol\beta_1 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol\beta_k \cdot \mathbf{X}_i}}$$

$$\Pr(Y_i = 2) = \frac{e^{\boldsymbol\beta_2 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol\beta_k \cdot \mathbf{X}_i}}$$

$$\cdots\cdots$$

$$\Pr(Y_i = K-1) = \frac{e^{\boldsymbol\beta_{K-1} \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol\beta_k \cdot \mathbf{X}_i}}$$

$$\Pr(Y_i = K) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol\beta_k \cdot \mathbf{X}_i}}$$

Se ajustan los K-1 conjuntos de coeficientes INDEPENDIENTEMENTE, sin embargo…..

Las porb. estan entre 0 y 1

y

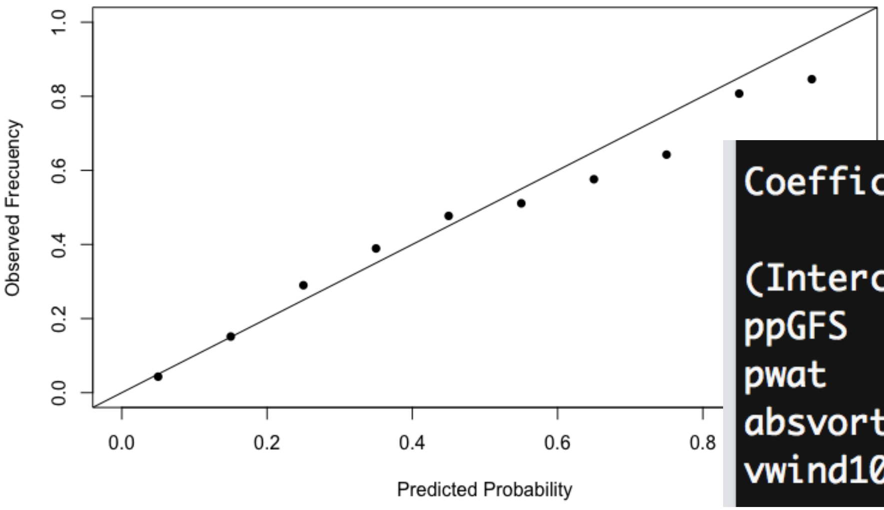
Las prob. suman 1

POR QUE ?????

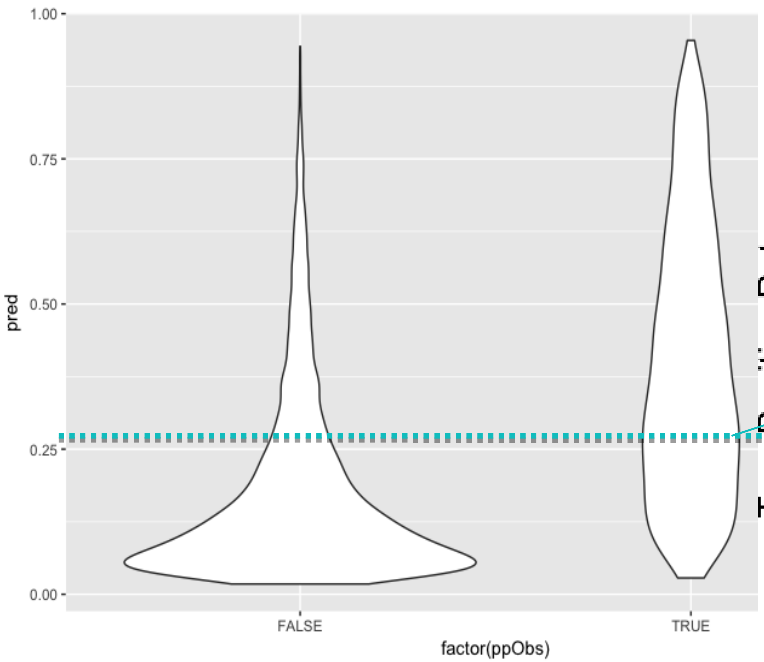
Clase de referencia

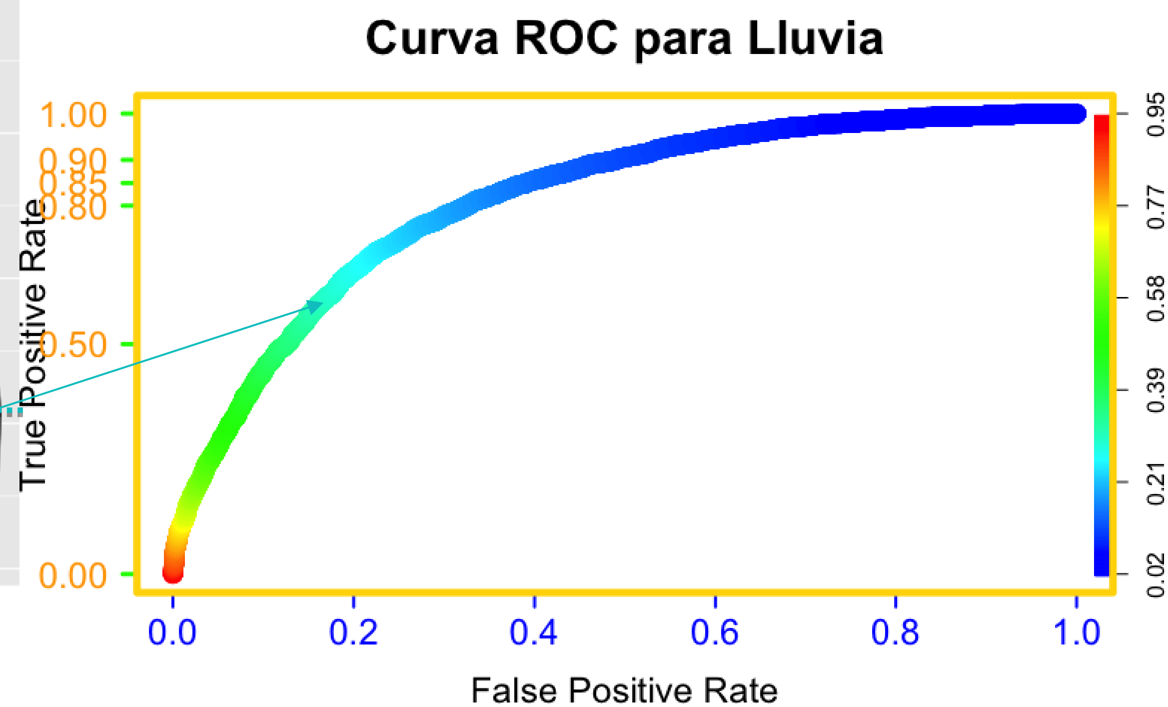
Función SOFTMAX

Ejemplo: Prediciendo Lluvia

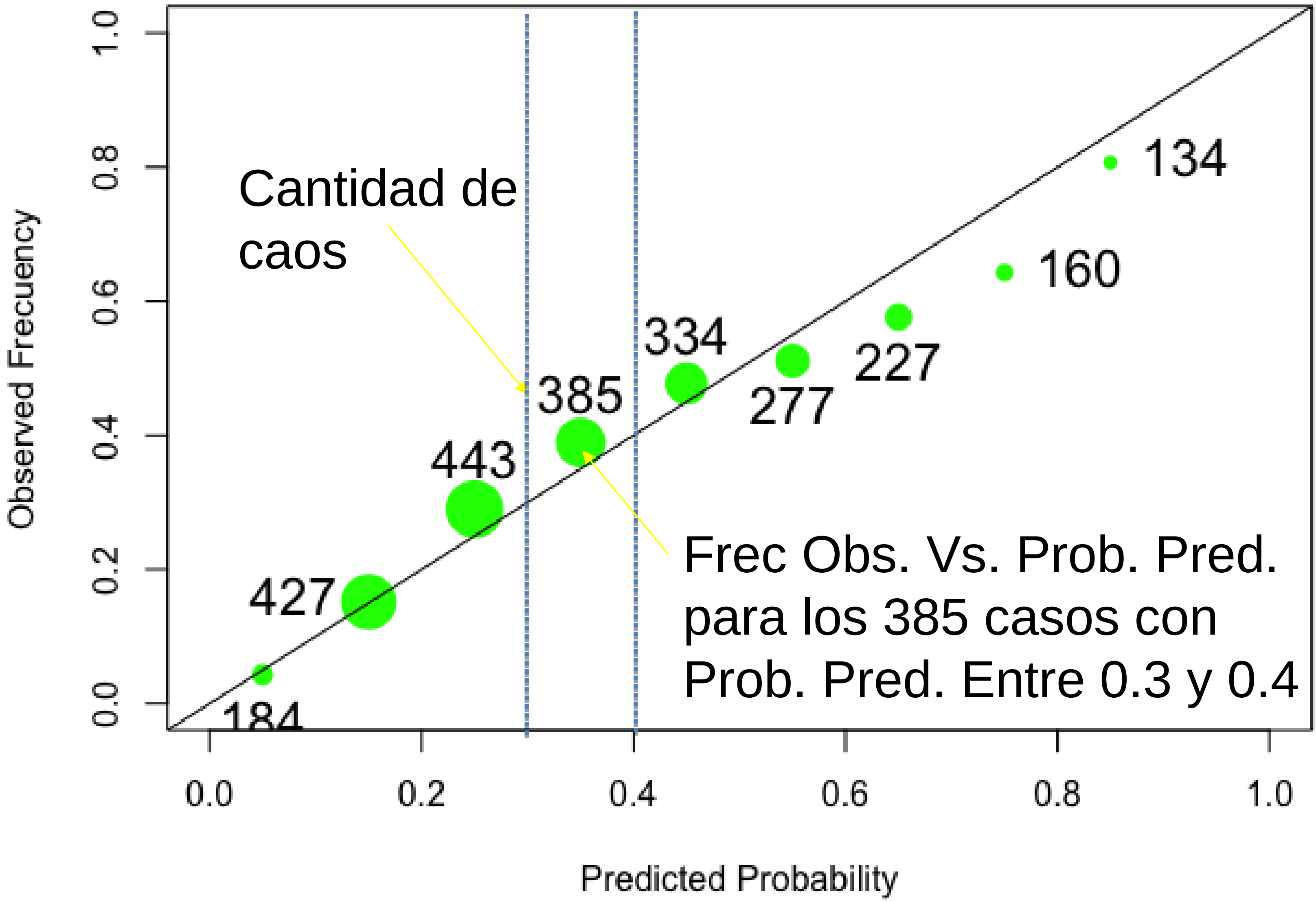



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.498e+00  9.565e-02 -47.023  < 2e-16 ***
ppGFS        2.226e-03  2.551e-03   0.873 0.382806
pwat         1.134e-01  2.963e-03  38.263  < 2e-16 ***
absvort     -3.839e+03  9.262e+02  -4.146 3.39e-05 ***
vwind1000   -2.387e-02  6.904e-03  -3.457 0.000546 ***
```

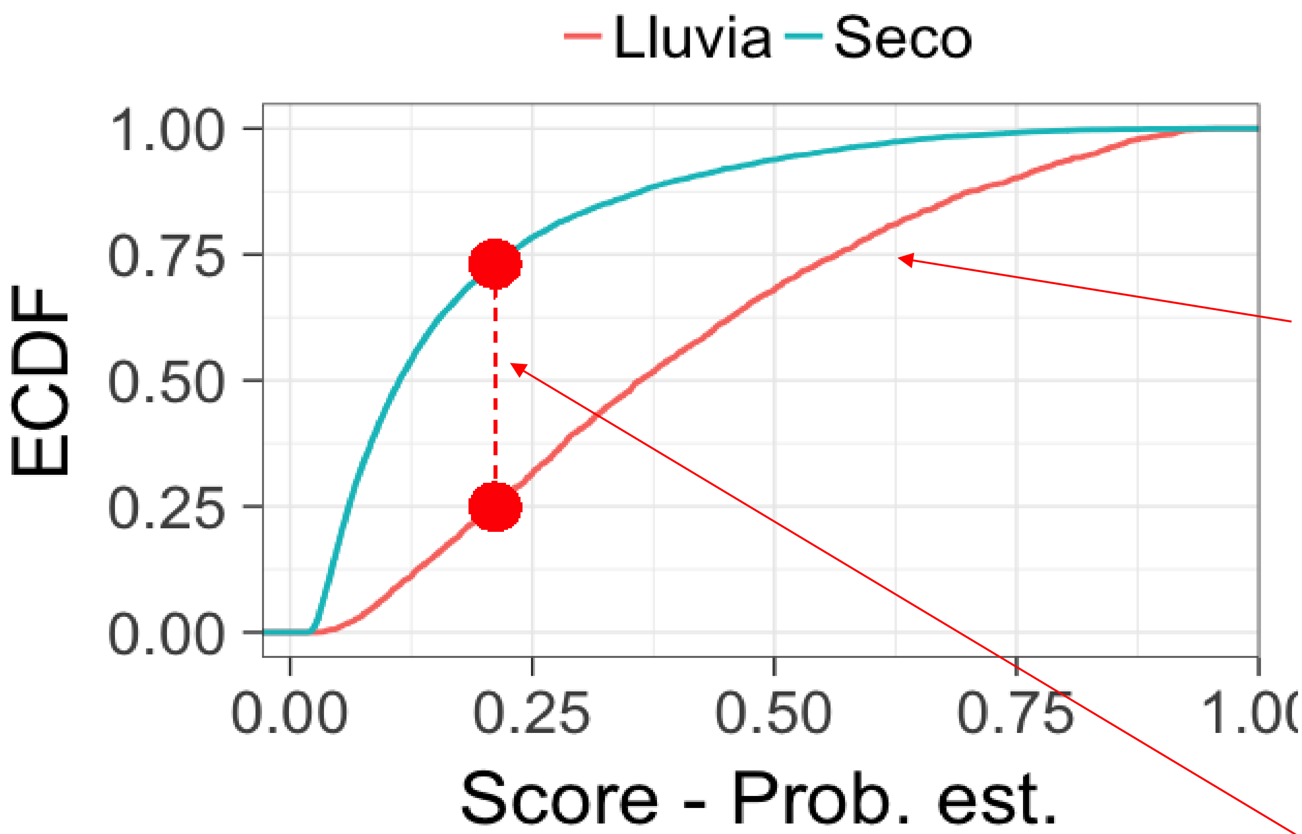




Verificando el Ajuste
Gráfico de Hosmer-Lemeshow

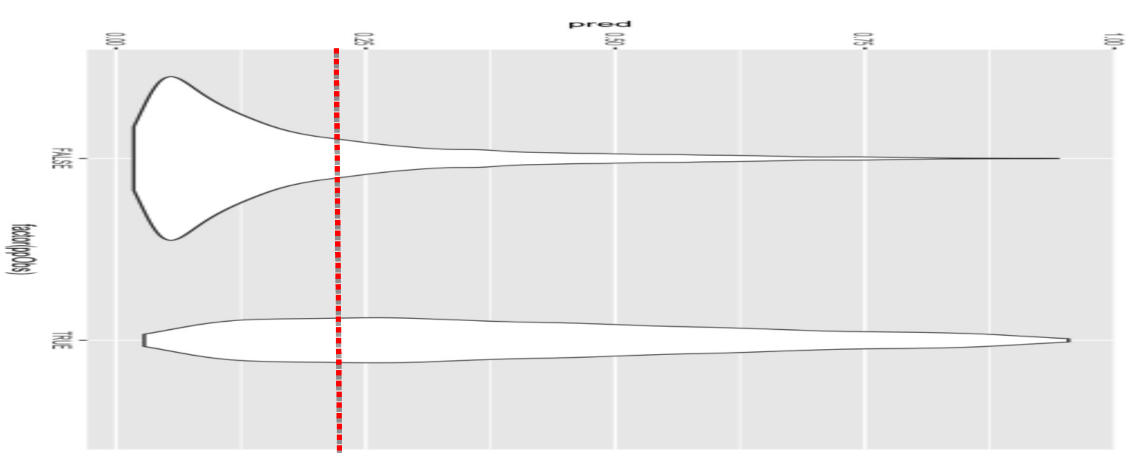


K-S Test: Seco / Lluvia



Función de Distribución Acumulada de los Scores de las Observaciones de Lluvia

Distancia KS



La Deviance como Residuos de un Modelo

Verosimilitud

Modelo Ajustado

$$Dev(Mod) = -2Log(\frac{L(Mod)}{L(Sat)}) \geq 0$$

Modelo Saturado

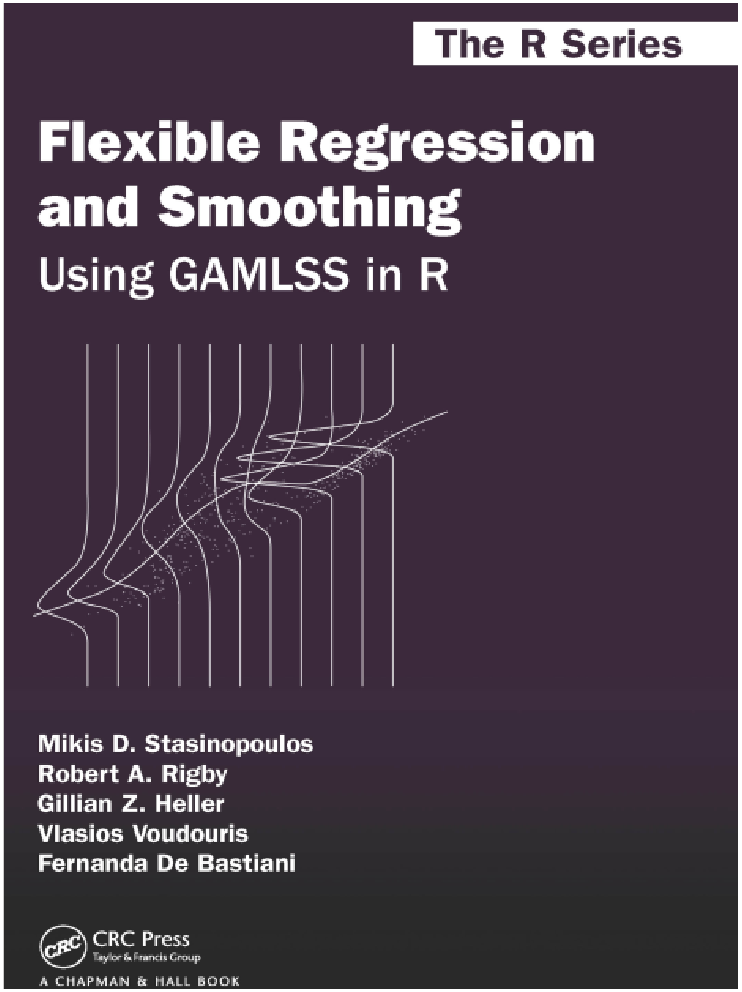
$$Deviance\ explicada = \frac{Dev(Mod_{NULL}) - Dev(Mod)}{Dev(Mod_{NULL})}$$

Modelo Nulo (solo con intercept)

Regresión Gamma

Técnica de Regresión

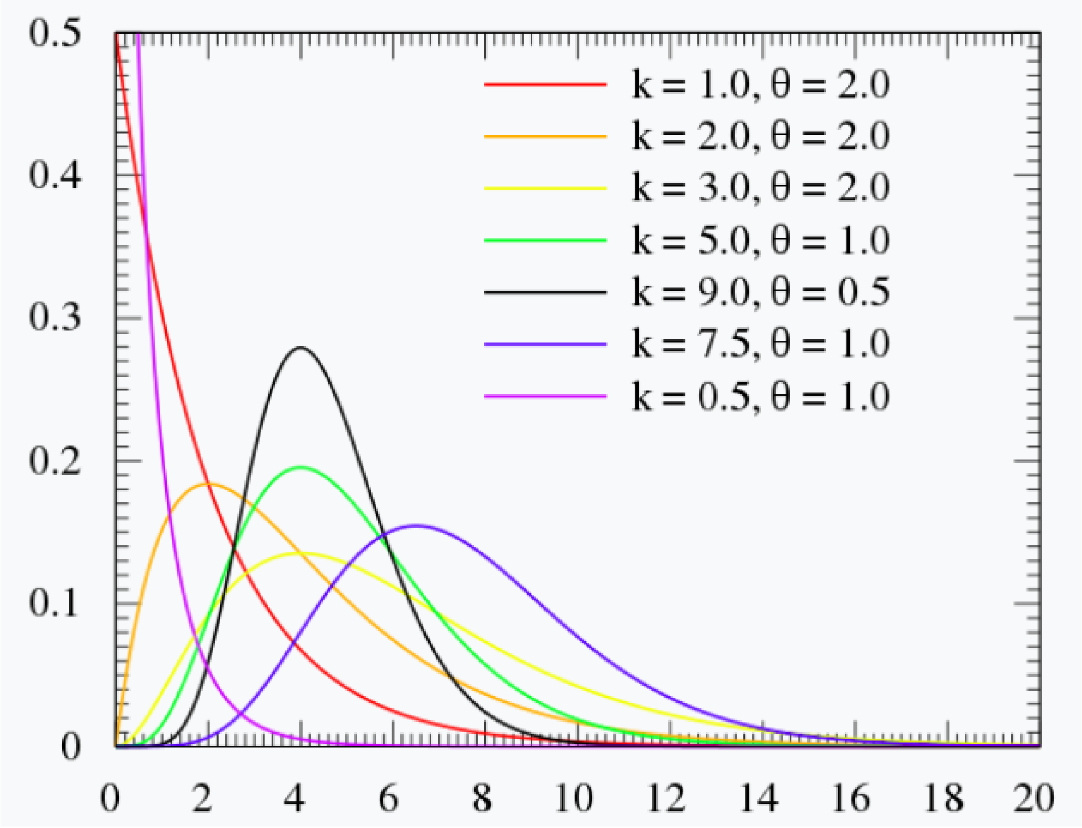
Bibliografía



Distribución Gamma

$$X \sim \Gamma(k, \theta) \equiv \mathrm{Gamma}(k, \theta)$$

$$f(x; k, \theta) = \frac{x^{k-1} e^{-\frac{x}{\theta}}}{\theta^k \Gamma(k)} \quad \text{for } x > 0 \text{ and } k, \theta > 0.$$



Parameters	• $k > 0$ shape • $\theta > 0$ scale
Support	$x \in (0, \infty)$
PDF	$\frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$
CDF	$\frac{1}{\Gamma(k)} \gamma\left(k, \frac{x}{\theta}\right)$
Mean	$\mathrm{E}[X] = k\theta$
Median	No simple closed form
Mode	$(k-1)\theta \text{ for } k \geq 1$
Variance	$\mathrm{Var}(X) = k\theta^2$

El Paquete GAMLSS

Parametrización standard con shape y scale

$$f(x;k,\theta) = \frac{x^{k-1}e^{-\frac{x}{\theta}}}{\theta^k\Gamma(k)} \quad \text{for } x > 0 \text{ and } k,\theta > 0.$$

Parametrización GAMLSS con Locación y Dispersión

$$f(y;\mu,\phi) = \frac{y^{1/\phi-1}\exp(-\frac{y}{\phi\mu})}{(\phi\mu)^{(1/\phi)}\Gamma(1/\phi)}, \qquad y>0,\ \mu>0,\ \phi>0.$$

$$\sigma = \sqrt{\phi}.$$

Dispersión

Link de Locación

$$\eta_1 = g_1(\mu) = \log(\mu)$$

$$\eta_2 = g_2(\sigma) = \log(\sigma)$$

Link de Dispersión

Predictores

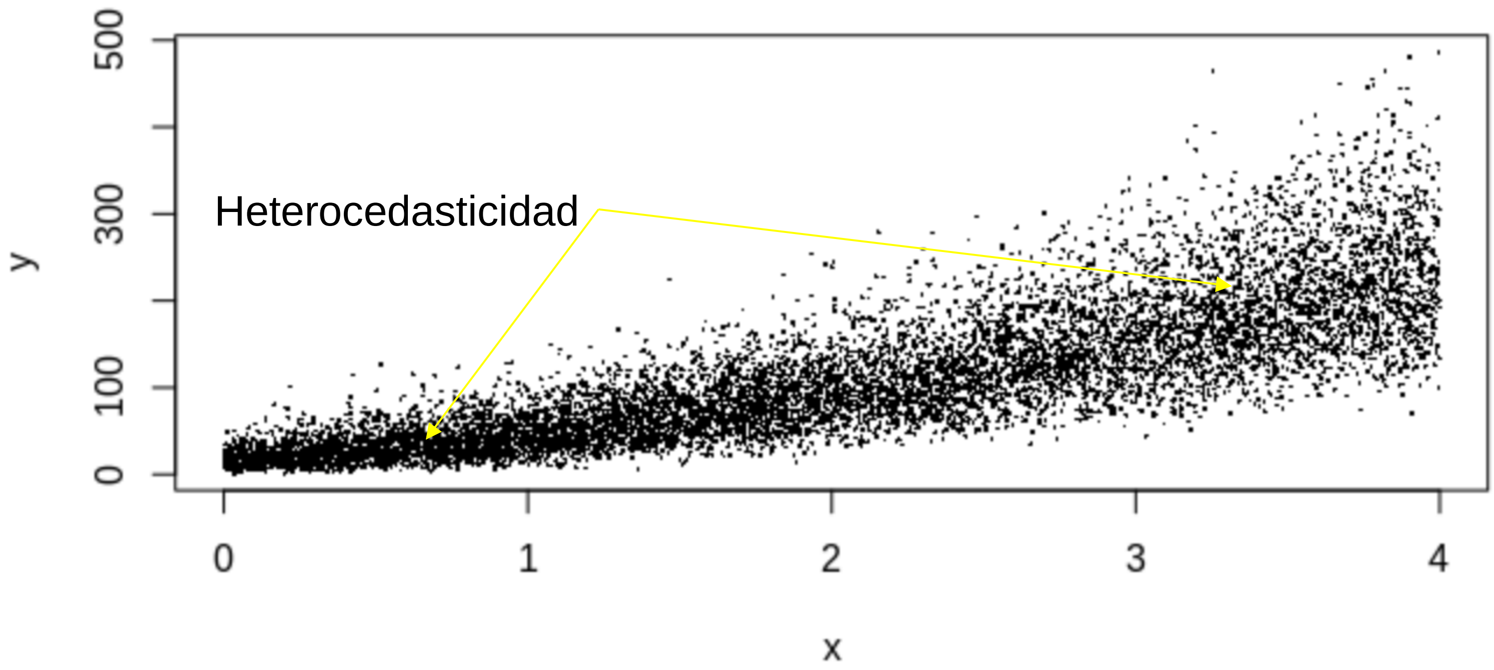
Lineales

$$\boldsymbol{\eta}_1 = \mathbf{X}\boldsymbol{\beta}_1$$

$$\boldsymbol{\eta}_2 = \mathbf{X}\boldsymbol{\beta}_2$$

Ejemplo Sencillo

```
n<-10000
x<-runif(n)*4
shape.vec<-3+2*x
scale.vec<-6+4*x
y<-rgamma(n=n,shape=shape.vec,scale=scale.vec)
```



El Ajuste

Splines con

Penalizaciones

(P-Splines)

```
fit_gamlss = gamlss(y ~ pb(x, df = 6),
                    data = test_gamma_df,
                    sigma.formula = ~pb(x, df = 6),
                    family = GA(mu.link = "log",
                    sigma.link ="log"))
```

Parametrización GAMLSS con Location y Scale

$$f(y;\mu,\phi) = \frac{y^{1/\phi-1}\exp(-\frac{y}{\phi\mu})}{(\phi\mu)^{(1/\phi)}\Gamma(1/\phi)}, \qquad y>0,\ \mu>0,\ \phi>0.$$

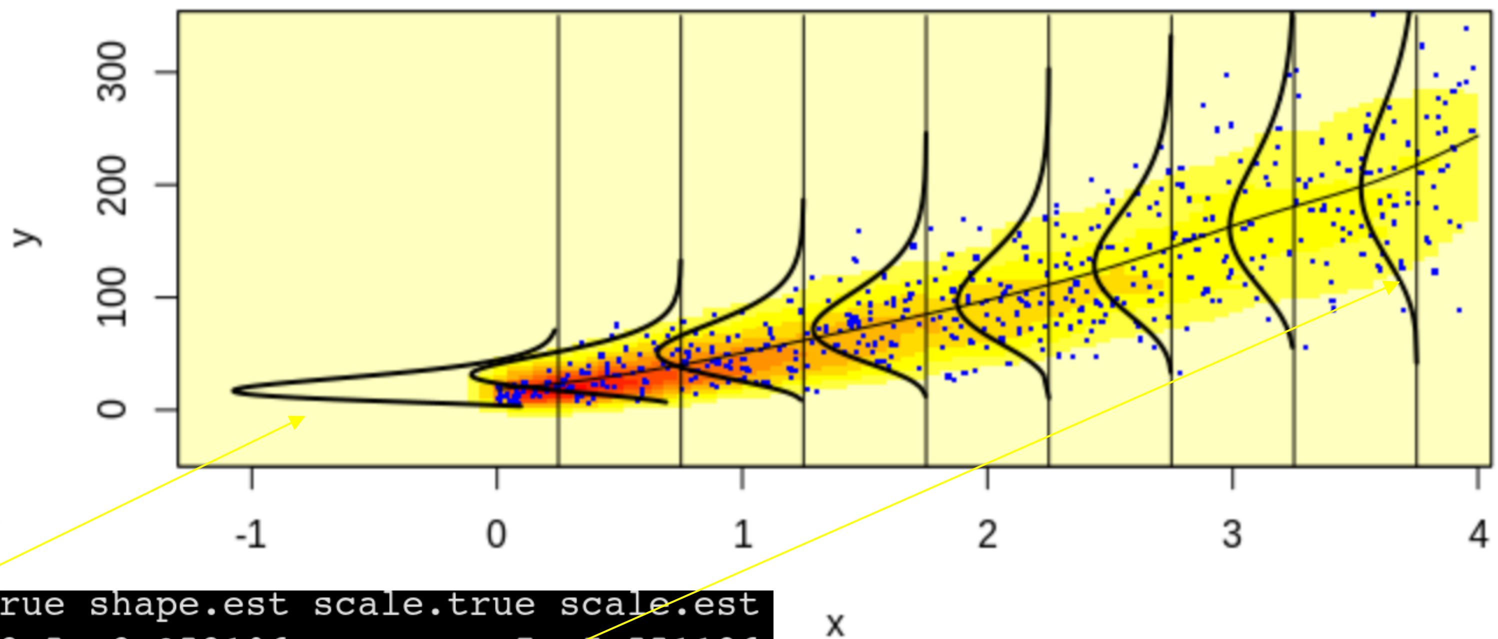
$$\sigma = \sqrt{\phi}.$$

$$\eta_1 = g_1(\mu) = \log(\mu) \qquad \boldsymbol{\eta}_1 = \mathbf{X}\boldsymbol{\beta}_1$$

$$\eta_2 = g_2(\sigma) = \log(\sigma) \qquad \boldsymbol{\eta}_2 = \mathbf{X}\boldsymbol{\beta}_2$$

Resultado del Ajuste

```
plotSimpleGamlss(y,x, model=fit_gamlss,
                          data=test_gamma_df,
                          x.val =secu,ylim=c(-50,350),xlim=c(-1.3,4))
```



	shape.true	shape.est	scale.true	scale.est
1	3.5	3.253106	7	7.551186
2	4.5	4.211414	9	9.343369
3	5.5	5.405384	11	11.145602
4	6.5	6.449820	13	13.065828
5	7.5	7.782119	15	14.197784
6	8.5	8.819463	17	16.380575
7	9.5	9.624951	19	18.883852
8	10.5	10.590221	21	21.082043

Modelos MUY Flexibles

Location

Dispertion

Shape

$$\mathbf{Y} \overset{\text{ind}}{\sim} \mathcal{D}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$$

$$\boldsymbol{\eta}_1 = g_1(\boldsymbol{\mu}) = \mathbf{X}_1\boldsymbol{\beta}_1 + s_{11}(\mathbf{x}_{11}) + \ldots + s_{1J_1}(\mathbf{x}_{1J_1})$$

$$\boldsymbol{\eta}_2 = g_2(\boldsymbol{\sigma}) = \mathbf{X}_2\boldsymbol{\beta}_2 + s_{21}(\mathbf{x}_{21}) + \ldots + s_{2J_2}(\mathbf{x}_{2J_2})$$

$$\boldsymbol{\eta}_3 = g_3(\boldsymbol{\nu}) = \mathbf{X}_3\boldsymbol{\beta}_3 + s_{31}(\mathbf{x}_{31}) + \ldots + s_{3J_3}(\mathbf{x}_{3J_3})$$

$$\boldsymbol{\eta}_4 = g_4(\boldsymbol{\tau}) = \mathbf{X}_4\boldsymbol{\beta}_4 + s_{41}(\mathbf{x}_{41}) + \ldots + s_{4J_4}(\mathbf{x}_{4J_4})$$

Smoothers

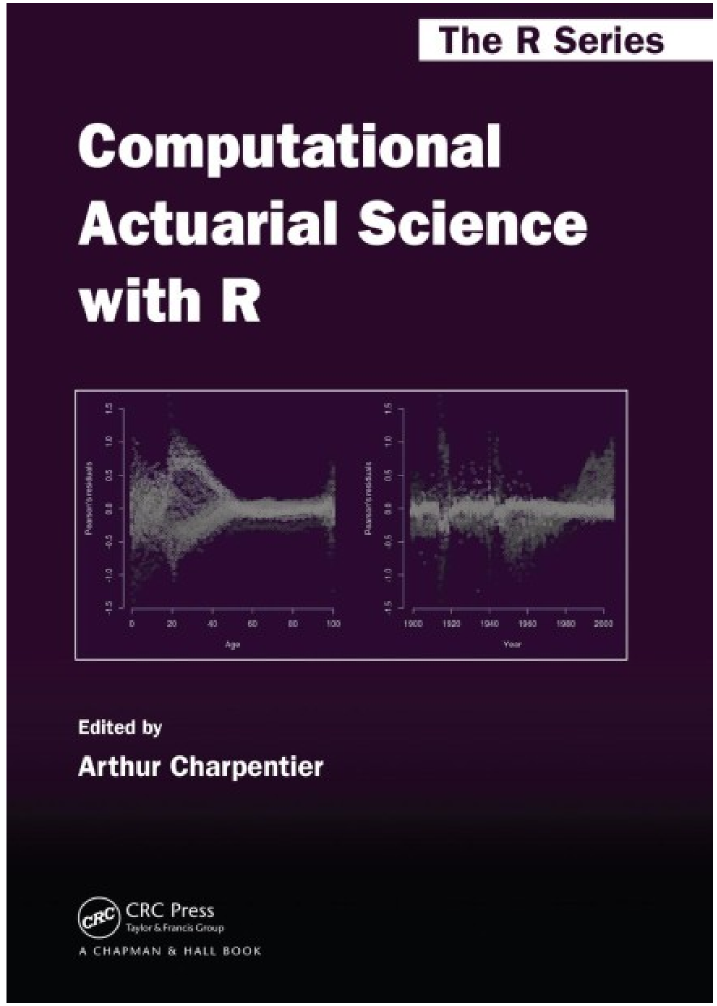
Modelo Lineal Generalizado (GLM) en la Ciencia Actuarial

Técnica de Modelado Paramétrico para Siniestralidad

¿ Para que usa GLM una Aseguradora?

- Para Predecir e Inferir la variable cantidad de siniestros (Frecuencia).
- Para Predecir e Inferir la variable magnitud del siniestro (Intensidad).
- Para Modelar la DISTRIBUCION de ambas variables, condicional a los factores diferenciadores del riesgo.

Bibliografía



¿ Como se Modela la Cantidad de Siniestros?

siniestros

del riesgo i

Tasa annual

(desconocida)

$$N_i \sim \mathcal{P}(\lambda)$$

siniestros

con una exposición (Ei)
distinta al año

Exposición del

riesgo i en años

$$Y_i \sim \mathcal{P}(\lambda \cdot E_i).$$

Verosimilitud de

todos los riesgos

$$\mathcal{L}(\lambda, \boldsymbol{Y}, \boldsymbol{E}) = \prod_{i=1}^{n} \frac{e^{-\lambda E_i} [\lambda E_i]^{Y_i}}{Y_i!},$$

Pero el Riesgo NO es Homogeneo !

Tasa annual específica

del riesgo i

siniestros

del riesgo i

$$\lambda_i = e^{\boldsymbol{X}'_i\boldsymbol{\beta}}$$

$$\boxed{Y_i \sim \mathcal{P}(E_i \cdot \lambda_i)}$$

Covariables

del riesgo i

La Exposición entra
como una variable mas,
con coef =1. Se lo llama

OFFSET

$$\boxed{Y_i \sim \mathcal{P}(e^{\boldsymbol{X}'_i\boldsymbol{\beta}+\log E_i})}$$

Ejemplo: Prediciendo La Cantidad de Siniestros en AUTOS

Exposición

siniestros

```
> str(dataCar)
'data.frame':   67856 obs. of  11 variables:
 $ veh_value: num  1.06 1.03 3.26 4.14 0.72 2.01 1.6 1.47 0.52 0.3
8 ...
 $ exposure : num  0.304 0.649 0.569 0.318 0.649 ...
 $ clm      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ numclaims: int  0 0 0 0 0 0 0 0 0 0 ...
 $ claimcst0: num  0 0 0 0 0 0 0 0 0 0 ...
 $ veh_body : Factor w/ 13 levels "BUS","CONVT",..: 4 4 13 11 4 5
8 4 4 4 ...
 $ veh_age  : int  3 2 2 2 4 3 3 2 4 4 ...
 $ gender   : Factor w/ 2 levels "F","M": 1 1 1 1 1 2 2 2 1 1 ...
 $ area     : Factor w/ 6 levels "A","B","C","D",..: 3 1 5 4 3 3 1
 2 1 2 ...
```

Resultado del Ajuste

```
v1<-veh_value
v2<-veh_value^2
v3<-veh_value^3
ajus<-glm(numclaims~gender*(v1+v2+v3)
          +offset(log(exposure)),family=poisson(link="log"))
```