
# Modelos lineales generalizados

## Máster en Data Science

**Mario Encinar, PhD, MAPFRE**
encinar@ucm.es

## El modelo de regresión lineal:

Se dispone de información de unos ciertos predictores $X = (X_1, \ldots, X_p)$ y una variable respuesta $Y$ que es continua. Se asume:

- Para $X = x$, $Y = x^t \cdot \beta + \epsilon$.
- $\epsilon \sim N(0, \sigma^2)$.
- $\epsilon$ y $X$ son independientes.

Naturalmente, en aprendizaje supervisado, todas las hipótesis en las que se basa el modelo de regresión lineal pueden relajarse.

**Posibles extensiones del modelo de regresión lineal:**

- La variable respuesta no tiene por qué ser continua.
- Los errores no tienen por qué seguir una distribución normal.
- No tiene por qué haber una relación lineal entre los predictores y la variable respuesta.

Los modelos lineales generalizados (GLM) sirven para extender el modelo de regresión lineal en todas estas direcciones de una forma *unificada*.

Supongamos que $Y$ es una variable aleatoria que sigue una cierta distribución de probabilidad dependiente de un parámetro $\theta$, cuya función de densidad (o, análogamente, función de masa) denotamos por $f = f(y; \theta)$.

Si $\theta$ es desconocido e $y_0$ es un valor observado de $Y$, la estimación más natural de $\theta$ es

$$\hat{\theta} = \arg\max_{\theta} f(y_0; \theta).$$

Cuando la función de densidad $f(y; \theta)$ se entiende de este modo, suele denotarse mediante $L(\theta; y)$ y llamarse la **función de verosimilitud** de $\theta$. Así, $\hat{\theta}$ es el **estimador de máxima verosimilitud** de $\theta$.

El método de máxima verosimilitud selecciona el conjunto de valores de los parámetros de la distribución que maximiza la función de verosimilitud. Intuitivamente elige los valores de los parámetros que hacen más probables los datos.

Véase:

- StatQuest: Maximum Likelihood, clearly explained!!!
- Why MLE?
- OLS-vs-MLE
- MLE-vs-OLS

*MLE: Maximum Likelihood Estimation/Estimator
*OLS: Ordinary Least Squares

Supongamos, por ejemplo, que $Y \sim \text{Pois}(\theta)$, con $\theta$ desconocido, y que $y_1, \ldots, y_n$ son $n$ observaciones independientes de $Y$. En tal caso, denotando $y = (y_1, \ldots, y_n)$, tenemos que la probabilidad de observar $y$ es

$$L(\theta; y) = f(y; \theta) = \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!}.$$

Maximizar $L(\theta; y)$ parece complicado, pero es equivalente a maximizar

$$\ell(\theta; y) = \log L(\theta; y) = \sum_{i=1}^{n} \left[-\theta + y_i \log(\theta) - \log(y_i!)\right],$$

que no lo es tanto.

$$\frac{d}{d\theta}\ell(\theta; y) = -n + \frac{1}{\theta}\sum_{i=1}^{n} y_i = 0 \Leftrightarrow \theta = \frac{1}{n}\sum_{i=1}^{n} y_i =: \hat{\theta}.$$

$$\left.\frac{d^2}{d\theta^2}\right|_{\theta=\hat{\theta}} \ell(\theta; y) = -\frac{1}{\hat{\theta}^2}\sum_{i=1}^{n} y_i < 0$$

De este modo,

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

es el estimador de maxima verosimilitud de $\theta$.

La técnica de estimación de máxima verosimilitud también se puede utilizar para ajustar modelos. Por ejemplo, para ajustar un modelo de regresión lineal como el de antes, donde suponemos que hay un único predictor $X$ y asumimos

- Para $X = x$, $Y = \beta_0 + \beta_1 x + \epsilon$.
- $\epsilon \sim N(0, \sigma^2)$.
- $\epsilon$ es independiente de $X$.

Bajo estas condiciones, si se observan $n$ patrones independientes $(x_i, y_i)$ y denotamos $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$, tenemos

$$f(y_i|x) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right) \exp\left(-\frac{1}{2}\left(\frac{y_i - (\beta_0 - \beta_1 x_i)}{\sigma}\right)^2\right),$$

¡Cuidado! Esto no implica que la variable dependiente (vector $y$) esté normalmente distribuída, sino que cada una de sus observaciones puede considerarse como una variable aleatoria normalmente distribuída cuya media es diferente, y depende de $x_i$. Véase: Does the assumption of Normal errors imply that Y is also Normal?

De lo anterior se puede obtener la función de verosimiltud:

$$L(\beta_0, \beta_1; x, y) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - (\beta_0 - \beta_1 x_i)}{\sigma}\right)^2\right),$$

donde hemos considerado, de nuevo, la probabilidad de que se den todas las observaciones como sucesos independientes.

Al tomar logaritmos:

$$\ell(\beta_0, \beta_1) = \log L(\beta_0, \beta_1; x, y) = -\frac{1}{n}\log(\sigma\sqrt{2\pi}) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - (\beta_0 - \beta_1 x_i)}{\sigma}\right)^2,$$

resulta que maximizar $\ell(\beta_0, \beta_1)$ devuelve el mismo resultado que la estimación por mínimos cuadrados ordinarios.

Supongamos que $Y$ es una variable aleatoria cuya función de densidad (o de masa) depende de un parámetro $\theta$. Decimos que $Y$ sigue una distribución de la familia exponencial si

$$f(y;\theta) = e^{a(y)b(\theta)+c(y)+d(\theta)},$$

donde $a$, $b$, $c$ y $d$ son funciones conocidas y suficientemente regulares.

Si hay más parámetros aparte de $\theta$, se tratan como si fueran constantes (contenidos en $a$ o $c$, digamos).

## Ejemplos de distribuciones de la familia exponencial:

- Normal.
- Exponencial.
- Gamma.
- Chi-cuadrado.
- Beta.
- Dirichlet.
- Bernoulli.
- Poisson.
- Binomial (con número de intentos fijo).
- Multinomial (con número de intentos fijo).

**La distribución normal (con desviación típica conocida) como miembro de la familia exponencial:**

$$
\begin{aligned}
f(y) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} = \exp\left(-\log(\sigma\sqrt{2\pi}) - \frac{(y-\mu)^2}{2\sigma^2}\right) \\
&= \exp\left(-\log(\sigma\sqrt{2\pi}) - \frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right) \\
&= \exp(yb(\theta) + c(y) + d(\theta))
\end{aligned}
$$

con $\theta = \frac{\mu}{\sigma^2}$, $b(\theta) = \theta$, $c(y) = -\frac{y^2}{2\sigma^2}$ y $d(\theta) = -\frac{\theta}{2\sigma} - \log(\sigma\sqrt{2\pi})$.

## La distribución Bernoulli como miembro de la familia exponencial:

$$
\begin{aligned}
f(y) &= \begin{cases} 1-p & \text{si} \quad y=0 \\ p & \text{si} \quad y=1 \end{cases} \\
&= \exp\left(y \log\left(\frac{p}{1-p}\right) + \log(1-p)\right), \quad y = 0, 1 \\
&= \exp(yb(\theta) + c(y) + d(\theta)), \quad y = 0, 1
\end{aligned}
$$

con $\theta = p$, $b(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$, $c(y) = 0$ y $d(\theta) = \log(1-\theta)$.

## Propiedades de las distribuciones de la familia exponencial:

Supuesto que

$$Y \sim f(y;\theta) = e^{yb(\theta)+c(y)+d(\theta)},$$

se tiene

$$E(Y) = -\frac{d'(\theta)}{b'(\theta)},$$

$$V(Y) = \frac{b''(\theta)d'(\theta) - d''(\theta)b'(\theta)}{b'(\theta)^3}$$

y, naturalmente,

$$\ell(\theta;y) = yb(\theta) + c(y) + d(\theta).$$

Véase:

- Univariate Distribution Relationships
- Introduction to the Bernoulli Distribution
- An Introduction to the Binomial Distribution
- An Introduction to the Poisson Distribution

Como hemos visto anteriormente, las distintas observaciones de la variable objetivo en un modelo de regresión lineal se modelizan mediante

$$Y_i \sim N(\mu_i, \sigma^2),$$

donde la relación con los predictores viene dada por

$$\mu_i = x_i^t \beta.$$

Los modelos lineales generalizados relajan ambas hipótesis.

## Componentes de un modelo generalizado lineal (GLM)

Un modelo lineal generalizado tiene tres componentes básicos:

- Componente aleatoria: Identifica la variable respuesta y su distribución de probabilidad.
- Componente sistemática: Especifica las variables explicativas (independientes o predictoras) utilizadas en la función predictora lineal.
- Función link: Es una función del valor esperado de $Y$ , $E(Y)$, como una combinación lineal de las variables predictoras.

## Componente aleatoria

La componente aleatoria de un GLM consiste en una variable aleatoria $Y$ con observaciones independientes $(y_1 \dots y_n)$.

En muchas aplicaciones, las observaciones de $Y$ son binarias, y se identifican como éxito y fracaso.

Aunque de modo más general, cada $y_i$ indicaría el número de éxitos de entre un número fijo de ensayos, y se modelizaría como una distribución binomial.

En otras ocasiones cada observación es un recuento, con lo que se puede asignar a $Y$ una distribución de Poisson o una distribución binomial negativa.

Finalmente, si las observaciones son continuas se puede asumir para $Y$ una distribución normal.

Todos estos modelos se pueden incluir dentro de la llamada familia exponencial de distribuciones.

## Componente sistemática

La componente sistemática de un GLM especifica las variables explicativas, que entran en forma de efectos fijos en un modelo lineal, es decir, las variables $x_j$ se relacionan como:

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Esta combinación lineal de variables explicativas se denomina predictor lineal. Alternativamente, se puede expresar como:

$$\eta_i := x_i^t \beta$$

con:

$$\beta^t = (\beta_0 \quad \beta_1 \ldots \beta_p)$$

y

$$x_{i,j=0} = 1$$

## Función link

Se denota el valor esperado de $Y$ como $\mu = E(Y)$; entonces, la función link especifica una función $g(\cdot)$ que relaciona $\mu$ con el predictor lineal como

$$g(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Así, la función link $g(\cdot)$ relaciona las componentes aleatoria y sistemática. De este modo, para $i = 1, \ldots, n$,

$$\mu_i = E(Y_i)$$

$$\eta_i = g(\mu_i) = x_i^t \beta$$

La función $g$ más simple es $g(\mu) = \mu$; esto es, la identidad que da lugar al modelo de regresión lineal clásico:

$$\mu = E(Y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Los modelos de regresión lineal típicos para respuestas continuas son un caso particular de los GLM.

**Hipótesis de un modelo lineal generalizado:** Dados $p$ predictores $X = (X_1, \ldots, X_p)$, una variable respuesta $Y$ y $n$ patrones observados $(x_i, y_i)$, $i = 1, \ldots, n$ (donde, para cada $i$, definimos $x_{i0} = 1$), asumimos que la relación entre $X$ e $Y$ es del siguiente modo: cada $y_i$ es un valor observado de una variable aleatoria $Y_i$, de modo que

1. Las $Y_i$ pertenecen a la familia exponencial, y tienen todas función de densidad (o de masa) de la forma

$$f(y_i; \theta_i) = e^{y_i b(\theta_i) + c(\theta_i) + d(y_i)}$$

   y media $\mu_i = \mu_i(\theta_i)$.
2. Las $Y_i$ son independientes entre sí.
3. La relación entre $X$ e $Y$ viene dada por

$$g(\mu_i) = x_i^t \beta,$$

   donde

$$\beta^t = (\beta_0 \;\; \beta_1 \ldots \beta_p)$$

   y $g$ es una función monótona y suficientemente regular, a la que se llama **función link**.

Estos modelos generalizan la regresión ordinaria de dos modos: permitiendo que $Y$ tenga distribuciones diferentes a la normal y, por otro lado, incluyendo distintas funciones link de la media. Esto resulta bastante útil para datos categóricos.

Los modelos GLM permiten la unificación de una amplia variedad de métodos estadísticos como la regresión, los modelos ANOVA y los modelos de datos categóricos. En realidad se usa el mismo algoritmo para obtener los estimadores de máxima verosimilitud en todos los casos. Este algoritmo es la base de la función glm de R.

**Ejemplo:** Consideramos un problema de clasificación binaria como, por ejemplo, determinar la supervivencia o no de los pasajeros del Titanic.

| | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked | Survived |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3 | Braund, Mr. Owen Harris | male | 22.00 | 1 | 0 | A/5 21171 | 7.2500 | | S | 0 |
| 2 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Thayer) | female | 38.00 | 1 | 0 | PC 17599 | 71.2833 | C85 | C | 1 |
| 3 | 3 | Heikkinen, Miss. Laina | female | 26.00 | 0 | 0 | STON/O2. 3101282 | 7.9250 | | S | 1 |
| 4 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.00 | 1 | 0 | 113803 | 53.1000 | C123 | S | 1 |
| 5 | 3 | Allen, Mr. William Henry | male | 35.00 | 0 | 0 | 373450 | 8.0500 | | S | 0 |
| 6 | 3 | Moran, Mr. James | male | *NA* | 0 | 0 | 330877 | 8.4583 | | Q | 0 |
| 7 | 1 | McCarthy, Mr. Timothy J | male | 54.00 | 0 | 0 | 17463 | 51.8625 | E46 | S | 0 |
| 8 | 3 | Palsson, Master. Gosta Leonard | male | 2.00 | 3 | 1 | 349909 | 21.0750 | | S | 0 |
| 9 | 3 | Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg) | female | 27.00 | 0 | 2 | 347742 | 11.1333 | | S | 1 |
| 10 | 2 | Nasser, Mrs. Nicholas (Adele Achem) | female | 14.00 | 1 | 0 | 237736 | 30.0708 | | C | 1 |
| 11 | 3 | Sandstrom, Miss. Marguerite Rut | female | 4.00 | 1 | 1 | PP 9549 | 16.7000 | G6 | S | 1 |
| 12 | 1 | Bonnell, Miss. Elizabeth | female | 58.00 | 0 | 0 | 113783 | 26.5500 | C103 | S | 1 |
| 13 | 3 | Saundercock, Mr. William Henry | male | 20.00 | 0 | 0 | A/5. 2151 | 8.0500 | | S | 0 |
| 14 | 3 | Andersson, Mr. Anders Johan | male | 39.00 | 1 | 5 | 347082 | 31.2750 | | S | 0 |
| 15 | 3 | Vestrom, Miss. Hulda Amanda Adolfina | female | 14.00 | 0 | 0 | 350406 | 7.8542 | | S | 0 |
| 16 | 2 | Hewlett, Mrs. (Mary D Kingcome) | female | 55.00 | 0 | 0 | 248706 | 16.0000 | | S | 1 |
| 17 | 3 | Rice, Master. Eugene | male | 2.00 | 4 | 1 | 382652 | 29.1250 | | Q | 0 |

**Primera tentativa: Modelo de probabilidad lineal** En respuestas binarias, un modelo análogo al de regresión lineal es

$$p(x) = \beta_0 + \beta x$$

que se denomina modelo de probabilidad lineal, ya que la probabilidad de éxito cambia linealmente con respecto a x.

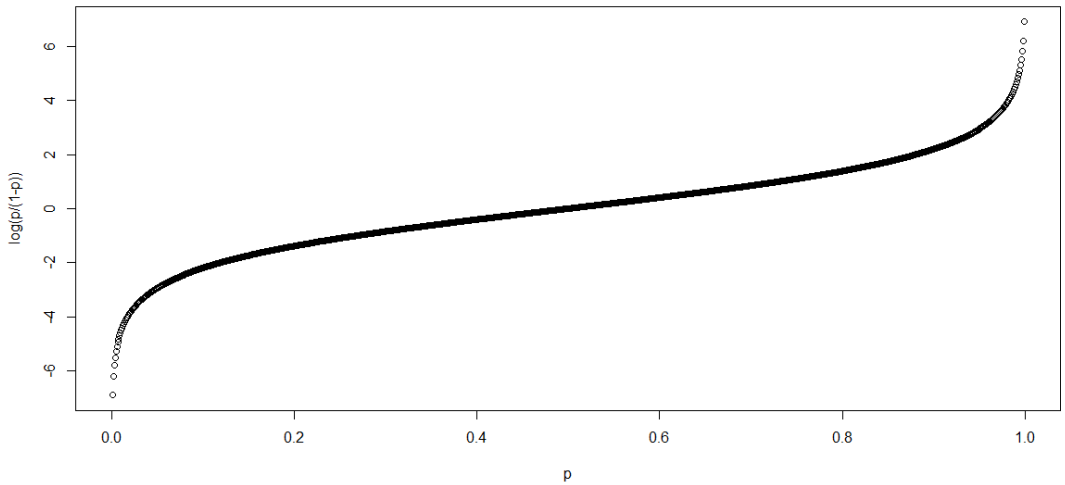
El parámetro $\beta$ representa el cambio en probabilidad por unidad de x.

Este modelo es un GLM con un componente aleatorio binomial y con función link igual a la identidad.

Sin embargo, este modelo tiene el problema de que aunque las probabilidades deben estar entre $0$ y $1$, el modelo puede predecir a veces valores $p(x) > 1$ y $p(x) < 0$.

**Podemos hacerlo mejor:** Modelizamos la supervivencia o no de cada individuo como una variable aleatoria $Y_i \sim \text{Ber}(p_i)$. $y_i = 0, 1$ son los valores observados de esta variable, y $p_i$ es el parámetro desconocido que queremos estimar (su probabilidad de supervivencia). La media de cada $Y_i$ es $\mu_i = p_i$.

Seguramente no tiene sentido hacer un modelo lineal del tipo $p_i \sim \ldots$, ya que $p_i \in [0, 1]$, pero la discusión anterior sugiere considerar la cantidad $\log\left(\frac{p_i}{1-p_i}\right)$



El modelo que se propone es

$$\log\left(\frac{p_i}{1-p_i}\right) = \sum_{j=0}^{p} \beta_j x_{ij},$$

donde $x_{i0} = 1$ y, para cada $i, j$, $x_{ij}$ es el valor observado del predictor $j$ para el individuo $i$.

A la función anterior se le denomina función **logit**.

Otra posibilidad de linealización del problema pasa por utilizar la función link **probit**, que mapea probabilidades $p_i \in [0, 1]$ en toda la recta real mediante $\Phi(p)$ donde $\Phi(\cdot)$ es la función cuantil de la distribución normal. Una alternativa también utilizada es la función link **log-log complementaria**.

Con esto, estamos bajo las asunciones del GLM que hemos comentado anteriormente.

## Ajuste de un GLM:

El vector de parámetros $\beta$ se estima por máxima verosimilitud. Razonando como en el caso del modelo de regresión lineal, se obtiene

$$\ell(\beta) = \sum_{i=1}^{n} [y_i b(\theta_i) + c(\theta_i) + d(y_i)]$$

y, para maximizar esto con respecto a $\beta$, debería resolverse el sistema de ecuaciones

$$\frac{\partial \ell}{\partial \beta_j} = 0, \ j = 0, 1, \ldots, p$$

y, posteriormente, comprobar que se verifican las condiciones de segundo orden (sobre la matriz hessiana de $\ell$). Esto, en general, no es sencillo (o factible) y, por eso

$$\hat{\beta} = \arg\max_{\beta} \ell(\beta)$$

suele estimarse mediante algún método numérico.

**Observación:** Nótese la dependencia de $\ell$ en $\beta$, a través de $\mu_i = \mu_i(\theta_i)$ y $g(\mu_i) = x_i^t \beta$.

## Regresión logística:

El modelo de regresión logística es el que hemos comentado en el ejemplo anterior: dado un problema de clasificación binaria, donde la variable objetivo toma valores $y_i = 0, 1$, se modeliza cada $Y_i$ como una Ber$(p_i)$ y se establece la relación

$$\log\left(\frac{p_i}{1-p_i}\right) = \sum_{j=0}^{p} \beta_j x_{ij}.$$

La cantidad $\frac{p_i}{1-p_i}$ son los *odds* en favor de que el individuo $i$ pertenezca a la clase "1", es decir, cuánto es más probable que $y_i = 1$ que $y_i = 0$.
Suele tomarse logaritmo de este cociente para linealizar, con lo que obtenemos las *log-odds* en favor de que el individuo $i$ pertenezca a la clase "1".

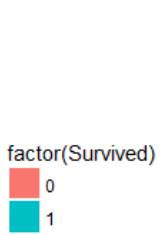
Expresando el modelo en forma *resumida*

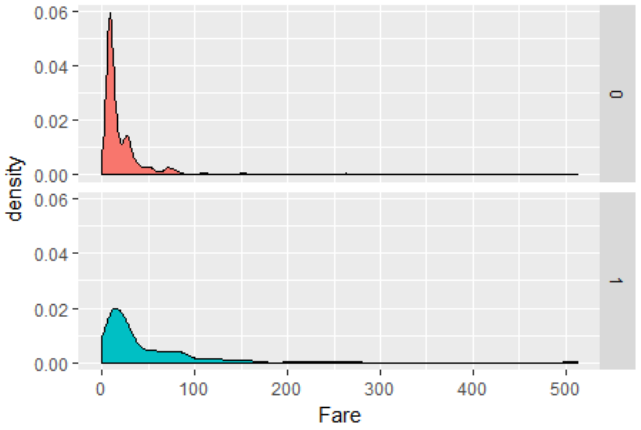
$$\log\left(\frac{p}{1-p}\right) = \sum_{j=0}^{p} \beta_j x_j .$$

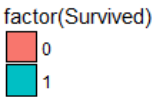
y asumiendo que todos los predictores son numéricos (las generalizaciones son inmediatas, al igual que en el caso del modelo de regresión lineal), podemos interpretar los coeficientes del modelo:

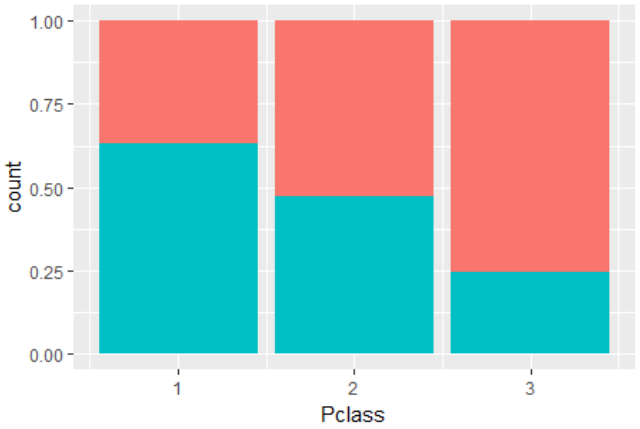
- $\beta_0 \sim \log \frac{P[Y=1|X=0]}{P[Y=0|X=0]}$.
- Para cada $j$, $\beta_j$ es el incremento que sufre $\log \frac{P[Y=1|X]}{P[Y=0|X]}$ cuando el predictor $X_j$ se incrementa en una unidad.

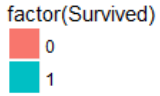

1.00
0.75
0.50
0.25
0.00
count
female
male
Sex

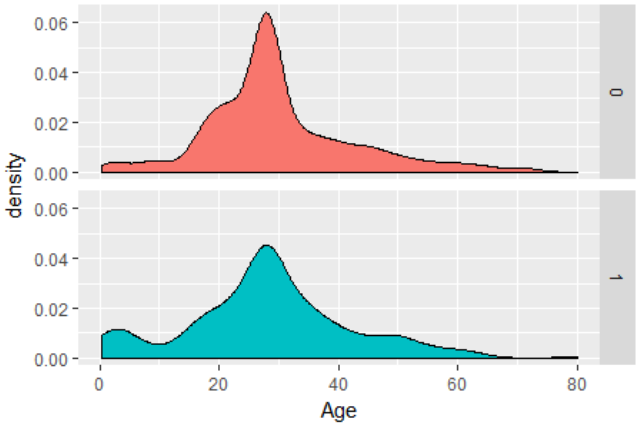

factor(Survived)
0
1

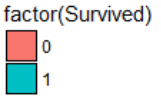

0.06
0.04
0.02
0.00
density
0
100
200
300
400
500
Fare
0
1


factor(Survived)
0
1


1.00
0.75
0.50
0.25
0.00
count
1
2
3
Pclass


factor(Survived)
0
1


0.06
0.04
0.02
0.00
density
0
20
40
60
80
Age
0
1


factor(Survived)
0
1

Ajustamos un modelo de regresión logística del tipo

$$\texttt{Survived} \sim \texttt{Sex, Pclass, Fare, Age}$$

y obtenemos el siguiente resultado:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.6553374  0.5085945   9.153  < 2e-16 ***
Sexmale     -2.6072959  0.1872514 -13.924  < 2e-16 ***
Pclass      -1.1529180  0.1355637  -8.505  < 2e-16 ***
Fare         0.0005922  0.0020347   0.291    0.771
Age         -0.0331244  0.0073991  -4.477 7.58e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1186.7  on 890  degrees of freedom
Residual deviance:  805.5  on 886  degrees of freedom
AIC: 815.5
```

Sobre la devianza, puede consultarse esto.

Las predicciones de nuestro modelo de regresión logística serán valores de probabilidad para cada una de las observaciones, $p_i$, que definen cada una de las binomiales que siguen, por hipótesis, las observaciones.
Estas predicciones se deben comparar con la variable objetivo, que es binaria (0: fracaso, 1: éxito).
Si en primera instancia consideramos que los valores $\hat{p_i} < 0.5$ reflejan "fracasos" y los $\hat{p_i} > 0.5$ se asocian con éxitos, tenemos, en nuestro ejemplo, la siguiente *matriz de confusión*:

```
> table(titanic$Survived, titanic$survived_prediction)

    FALSE TRUE
0   433  116
1    72  270
```

que permite evaluar la bondad del modelo.

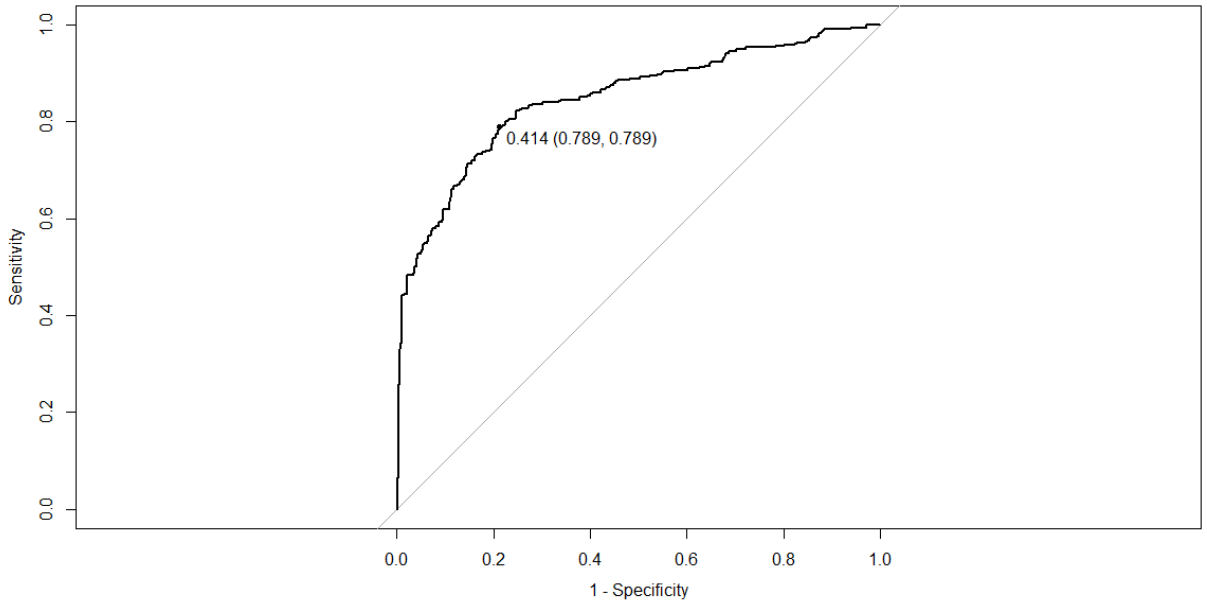
| | | *Predicted class* | | |
|---|---|---|---|---|
| | | − or Null | + or Non-null | Total |
| *True class* | − or Null | True Neg. (TN) | False Pos. (FP) | N |
| | + or Non-null | False Neg. (FN) | True Pos. (TP) | P |
| | Total | $N^*$ | $P^*$ | |

| Name | Definition | Synonyms |
|---|---|---|
| False Pos. rate | FP/N | Type I error, 1−Specificity |
| True Pos. rate | TP/P | 1−Type II error, power, sensitivity, recall |
| Pos. Pred. value | $TP/P^*$ | Precision, 1−false discovery proportion |
| Neg. Pred. value | $TN/N^*$ | |

Anteriormente hemos elegido $0.5$ como umbral de discriminación entre éxitos y fracasos pero si lo cambiamos $p_{threshold}$ en la siguiente ecuación:

$$\hat{p_i} < p_{threshold}$$

podemos construir la curva ROC:




## Understanding ROC curves

## Regresión logística *sólo* con predictores categóricos:

Supongamos que al ajustar un modelo de regresión logística todos los predictores son categóricos. Entonces, la información contenida en el set de datos puede resumirse contando, para cada combinación posible de niveles de los predictores, cuántos elementos pertenecen a cada clase (según variable objetivo).

| Sex | FClass | Num | Num_Survived |
|---|---|---|---|
| female | FALSE | 220 | 142 |
| female | TRUE | 94 | 91 |
| male | FALSE | 455 | 64 |
| male | TRUE | 122 | 45 |

En este caso, podemos modelizar el número de “unos” para cada combinación posible de niveles de los predictores como

$$Y_i \sim \text{Bi}(n_i, p_i),$$

donde $n_i$ es el número de individuos para los que se da esa combinación en los predictores.

Como $E[Y_i] = n_i p_i$, tiene sentido ajustar un GLM de la forma

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \ldots \beta_p x_{ip}.$$

Es fácil ver que este modelo será equivalente al que se habría obtenido con los datos desagregados.

## Regresión logística para predicción de variables categóricas no-binarias:

Si la variable objetivo $Z$ es una variable categórica que puede tomar $J$ posibles valores $z_1, \ldots, z_J$, definimos

$$Y_1 = I[Z = z_1], \ldots, Y_{J-1} = I[Z = z_{J-1}].$$

Entonces, se pueden ajustar $J - 1$ modelos de regresión logística para estimar

$$P[Z = z_1] = P[Y_1 = 1], \ldots, P[Z = z_{J-1}] = P[Y_{J-1} = 1]$$

y, naturalmente

$$P[Z = z_J] = 1 - \sum_{j=1}^{J-1} P[Z = z_j].$$

## Modelos de Poisson:

La distribución de Poisson es una distribución de probabilidad discreta que expresa, a partir de una frecuencia de ocurrencia media, la probabilidad de que ocurra un determinado número de eventos durante un cierto periodo de tiempo.

- Número de erratas por página que escribe una cierta persona.
- Número de e-mails recibidos al día por una cierta persona.
- Número de siniestros sufridos por un coche al año.
- Número de premios recibidos por estudiantes de un cierto instituto en un curso académico.

Su función de masa viene dada por

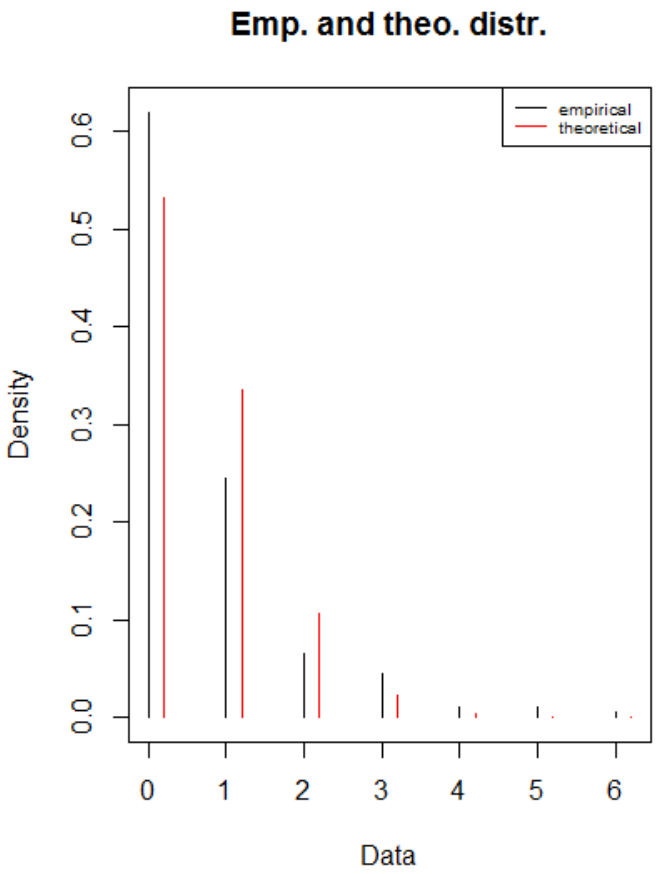
$$P[Y = y] = e^{-\lambda}\frac{\lambda^y}{y!} = e^{-\lambda + y\log(\lambda) - \log(y!)},\ y = 0, 1, 2\ldots$$

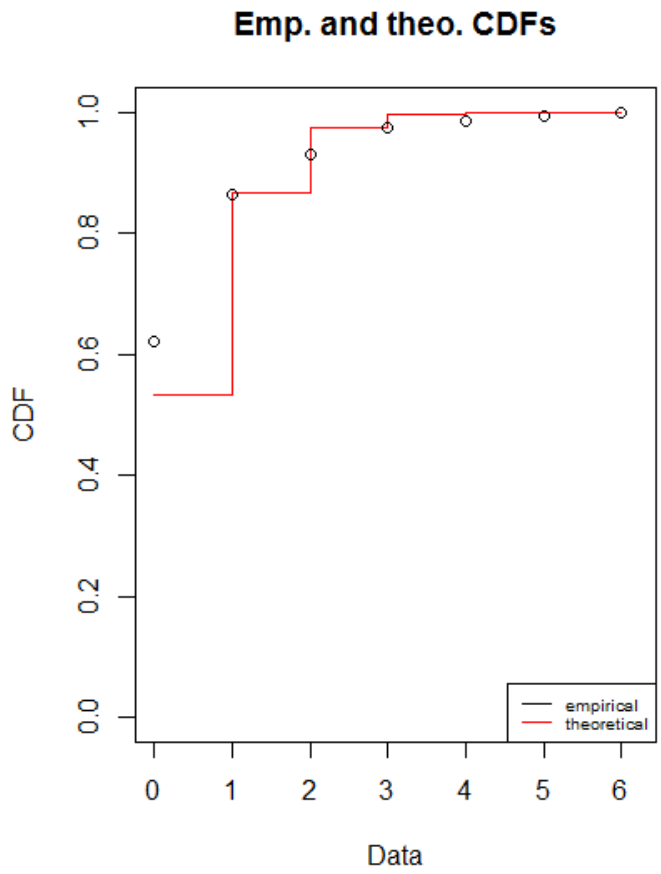
donde $0 < \lambda = E[Y] = V[Y]$ es el número de veces medio que se espera que suceda el evento en tal periodo de tiempo.

El dataset `awards.csv` (ver esto) contiene información sobre el número de premios que recibieron durante un curso académico los estudiantes de un cierto instituto, junto con información acerca de su nota en Matemáticas y el tipo de programa en el que estaban inscritos.

```
   num_awards          prog          math      
 Min.   :0.00   General    : 45   Min.   :33.0  
 1st Qu.:0.00   Academic   :105   1st Qu.:45.0  
 Median :0.00   Vocational: 50   Median :52.0  
 Mean   :0.63                    Mean   :52.6  
 3rd Qu.:1.00                    3rd Qu.:59.0  
 Max.   :6.00                    Max.   :75.0  
```

¿Tiene sentido modelizar `num_awards` como una distribución de Poisson?

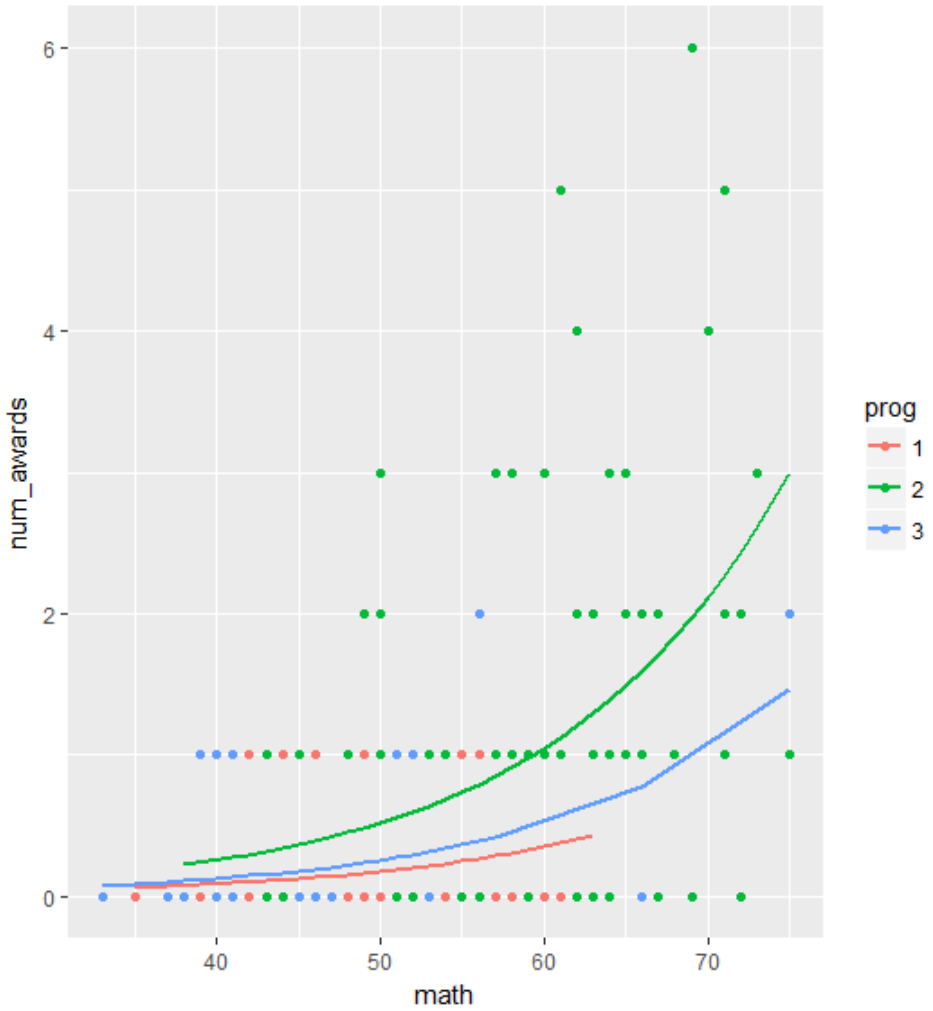


Dado que $\lambda > 0$, seguramente para ajustar un GLM de tipo Poisson es más natural considerar como función link el logaritmo, de modo que el modelo resultante es de la forma

$$\log(\lambda) = \sum_{i=0}^{p} \beta_j x_j.$$

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.24712    0.65845  -7.969 1.60e-15 ***
prog2        1.08386    0.35825   3.025  0.00248 **
prog3        0.36981    0.44107   0.838  0.40179
math         0.07015    0.01060   6.619 3.63e-11 ***
```


num_awards
6
4
2
0
40
50
60
70
math
prog
1
2
3

## Análisis de supervivencia:

La distribución exponencial es una distribución de probabilidad continua que sigue el *tiempo transcurrido* hasta que sucede un determinado evento.

- Tiempo transcurrido en un *call-center* hasta que se recibe la primera llamada del día.
- Cantidad de metros de hilo en una bobina que hay que desenrollar hasta encontrar un nudo.
- Tiempo que transcurre hasta que un cierto teléfono tiene su primera avería.

Su función de densidad viene dada por

$$f(y) = \lambda e^{-\lambda y} = e^{\log(\lambda) - \lambda y},\ y > 0.$$

$\lambda > 0$ es el inverso del tiempo medio de espera hasta que sucede el evento. Se tiene $E[Y] = \lambda^{-1}$, $V[Y] = \lambda^{-2}$.

La función link que se utiliza generalmente es $g(\theta) = \log(\theta)$.

Otra distribución que se utiliza con frecuencia para modelizar este tipo de fenómenos es la Weibull (ver esto).

## Modelos gamma:

La distribución Gamma es una distribución de probabilidad continua que suele emplearse para modelizar variables positivas, asimétricas y con colas pesadas.

- Costes.
- Tiempos de espera.

Su función de densidad viene dada por

$$f(y) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y},\ y > 0.$$

Se tienen $E[Y] = \frac{\alpha}{\beta}$ y $V[Y] = \frac{\alpha}{\beta^2}$.

Las funciones link que se utilizan generalmente son $g(\theta) = \frac{1}{\theta}$, $g(\theta) = \theta$ y $g(\theta) = \log(\theta)$.

1. Plantear el problema en forma de aprendizaje supervisado.
2. Identificar (intuir) qué distribución sigue la variable respuesta.
3. Ajustar un GLM:

**Common distributions with typical uses and canonical link functions**

| Distribution | Support of distribution | Typical uses | Link name | Link function | Mean function |
|---|---|---|---|---|---|
| Normal | real: $(-\infty, +\infty)$ | Linear-response data | Identity | $\mathbf{X}\boldsymbol{\beta} = \mu$ | $\mu = \mathbf{X}\boldsymbol{\beta}$ |
| Exponential | real: $(0, +\infty)$ | Exponential-response data, scale parameters | Inverse | $\mathbf{X}\boldsymbol{\beta} = \mu^{-1}$ | $\mu = (\mathbf{X}\boldsymbol{\beta})^{-1}$ |
| Gamma | | | | | |
| Inverse Gaussian | real: $(0, +\infty)$ | | Inverse squared | $\mathbf{X}\boldsymbol{\beta} = \mu^{-2}$ | $\mu = (\mathbf{X}\boldsymbol{\beta})^{-1/2}$ |
| Poisson | integer: $0, 1, 2, \ldots$ | count of occurrences in fixed amount of time/space | Log | $\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$ | $\mu = \exp(\mathbf{X}\boldsymbol{\beta})$ |
| Bernoulli | integer: $\{0, 1\}$ | outcome of single yes/no occurrence | Logit | $\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$ | $\mu = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})} = \frac{1}{1+\exp(-\mathbf{X}\boldsymbol{\beta})}$ |
| Binomial | integer: $0, 1, \ldots, N$ | count of # of "yes" occurrences out of N yes/no occurrences | | | |
| Categorical | integer: $[0, K)$ | outcome of single K-way occurrence | | | |
| | K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1 | | | | |
| Multinomial | *K*-vector of integer: $[0, N]$ | count of occurrences of different types (1 .. *K*) out of *N* total *K*-way occurrences | | | |

(Fuente: esto.)

4. Validar, optimizar, extraer conclusiones...

En el campo de los seguros de no-vida, es crítico para la aseguradora estimar de forma correcta la **prima pura** asociada a cada póliza, es decir, el importe que necesita percibir para asumir las consecuencias de los riesgos que le son transferidos.

Una forma sencilla y conservadora de estimar la prima pura es igualarla al gasto medio por póliza del año anterior, pero esto no es *competitivo*.

Lo ideal sería predecir los costes que va a causar cada póliza.

El modelado general es el siguiente:

$$\text{Coste} = \text{Número de siniestros} \times \text{Coste medio por siniestro.}$$

- El número de siniestros se modela mediante un GLM de tipo Poisson.
- El coste medio por siniestro se modela mediante un GLM de tipo Gamma.
- La *credibilidad* que se otorga a cada una de las observaciones se pondera en base a su exposición (la fracción del año que ha transcurrido desde que fue contratada la póliza).

A Practitioner's Guide to Generalized Linear Models

Guller, D., Goldburg, M.. *The Kaggle Challenge*

**Referencias:**


- Dobson, A., Barnett, A. (2008). *An Introduction to Generalized Linear Models* (Third Edition). CRC Press. ISBN: 978-1-58488-950-2. Disponible aquí
- Fox, J. (2002). *Linear Mixed Models. Appendix to An R and S-Plus Companion to Applied Regression*. Disponible aquí.
- Hastie, T., Tibshirani, R., Friedman, J. (2009). *The elements of Statistical Learning. Data Mining, Inference and Prediction.* Springer. ISBN: 978-0-387-84858-7.
- James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R.* Springer. ISBN: 978-1-4614-7138-7.
- Nelder, J.A., Wedderburn, R.W.M. (1972). *Generalized Linear Models*, Journal of the Royal Statistical Society. Series A (General), Vol. 135, No. 3, pp. 370-384.
- Ohlsson, E., Johansson, B. (2010). *Non-life insurance pricing with Generalized Linear Models.* Springer. ISBN: 978-3-642-10790-0.