Aplicaciones de modelos lineales generalizados

Los modelos lineales generalizados son una familia de modelos para el análisis estadístico. Incluye la regresión lineal y logística como casos especiales.

Un modelo lineal generalizado consiste de:

1. Un vector de datos $y = (y_1, \ldots, y_n)$
2. Predictores X y coeficientes β para construir un predictor lineal $X\beta$.
3. Una función de enlace g que da como resultado datos transformados

$$E(Y) = g^{-1}(X\beta)$$

   que son usados para modelar los datos.
4. Una distribución para los datos $p(y|X)$.
5. Posiblemente otros parámetros, como varianza, o puntos de corte, involucrados en los predictores, o bien, la función enlace o la distribución de los datos.

La regresión lineal predice directamente datos continuos y de un predictor lineal $X\beta = \beta_0 + X_1\beta_1 + \cdots + X_k\beta_k$.

`Otros modelos que vamos a ver son:`

1. El modelo *logístico-binomial* se utiliza en casos cuando los datos observados y_i representan el número de éxitos en n_i ensayos independientes. En este modelo la función liga es logit y la distribución de los datos es binomial. Al igual que con la regresión Poisson, el modelo binomial típicamente se puede mejorar agregando un parámetro de sobredispersión.

2. El modelo *probit* es igual que regresión logística pero se reemplaza la función liga por la *distribución normal acumulada*. Se puede pensar como usar la distribución normal en los errores estimados del modelo.

3. El *modelo Poisson* se utiliza para datos de *conteos*; es decir, donde cada dato observado y_i puede ser igual a $0, 1, 2, \ldots$. La función liga que se utiliza habitualmente g es logarítmica, de modo que $g(x) = \exp(x)$ transforma un predictor lineal continuo $X_i\beta$ en un y_i positivo. La distribución de datos es Poisson. A veces es buena idea agregar un parámetro a este modelo para capturar la **sobredispersión**, es decir, la variación en los datos más allá de la que captura el modelo.

Función glm()

glm(variable respuesta $\sim$ variable explicativa1 + ...+variable explicativa p, family = familia(link="función de enlace"), data = datos)

Familia Función de enlace por defecto:

binomial (link = "logit")

gaussian (link = "identity")

Gamma (link = "inverse")

inverse.gaussian (link = "1/mu^2")

poisson (link = "log")

quasi (link = "identity", variance = "constant")

quasibinomial (link = "logit")

quasipoisson (link = "log")

Regresión Binomial: Datos binarios

La regresión logística predice $P(y = 1)$ para datos binarios a partir de un predictor lineal transformado por la función logística inversa.

En el siguiente ejemplo se modela la probabilidad de fraude por impago (default = Y) en función del balance de la cuenta bancaria (balance = X).

Ejemplo 1: Interpretación de coeficientes de regresión

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.0     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ISLR)
datos <- Default

head(datos)
```

```
##   default student   balance    income
## 1      No      No  729.5265 44361.625
## 2      No     Yes  817.1804 12106.135
## 3      No      No 1073.5492 31767.139
## 4      No      No  529.2506 35704.494
## 5      No      No  785.6559 38463.496
## 6      No     Yes  919.5885  7491.559
```

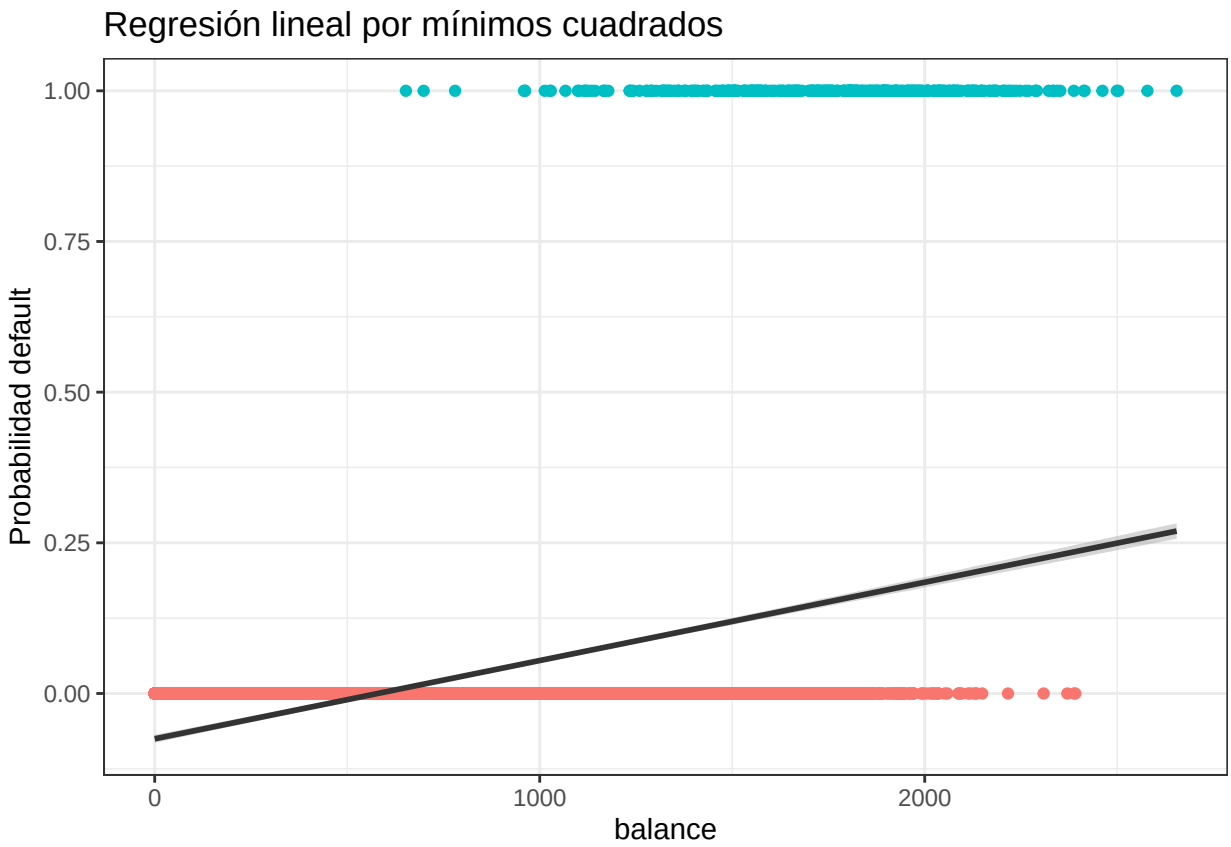
```
# Se recodifican los niveles No, Yes a 0 y 1
datos <- datos %>%
  select(default, balance) %>%
  mutate(default = recode(default,
                          "No"  = 0,
                          "Yes" = 1))
head(datos)
```

```
##   default   balance
## 1       0  729.5265
## 2       0  817.1804
## 3       0 1073.5492
## 4       0  529.2506
## 5       0  785.6559
## 6       0  919.5885
```

```
# Modelo de regresión Lineal
modelo_lineal <- lm(default ~ balance, data = datos)

# Representación gráfica del modelo.
```

```
ggplot(data = datos, aes(x = balance, y = default)) +
  geom_point(aes(color = as.factor(default))) +
  geom_smooth(method = "lm", color = "gray20") +
  theme_bw()  +
  labs(title = "Regresión lineal por mínimos cuadrados",
       y = "Probabilidad default") +
  theme(legend.position = "none")
```




```
predict(object = modelo_lineal, newdata = data.frame(balance = 100))
```

```
##           1
## -0.06220474
```

Modelo mal planteado, se predice la probabilidad de default para alguien que tiene un balance de 100, el valor obtenido es menor que 0.

Modelo de regresión Logística binaria

La asociación entre X e Y es no lineal, el modelo de regresión logística asocia la variable explicativa X con la media de Y a través de una función de enlace.

$$E(y_i) = p_i$$

$$logit(p_i) = \beta_0 + \beta_1 x_i$$

$$p(Y = 1|X = x) = p_i = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x_i}}$$

Puede interpretarse como: la probabilidad de que la variable cualitativa Y adquiera el valor 1 (el nivel de referencia, codificado como 1, probabilidad de éxito), dado que el predictor X_i tiene el valor x_i.

A partir de dicha definición se tiene que:

$$logit(p_i) = \beta_0 + \beta_1 x_i.$$

Ajuste de un modelo logístico usando la función glm():

```
modelo_logistico <- glm(default ~ balance, data = datos, family = "binomial")

summary(modelo_logistico)
```

```
##
## Call:
## glm(formula = default ~ balance, family = "binomial", data = datos)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2697  -0.1465  -0.0589  -0.0221   3.7589
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
## balance      5.499e-03  2.204e-04   24.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1596.5  on 9998  degrees of freedom
## AIC: 1600.5
##
## Number of Fisher Scoring iterations: 8
```

```
modelo_logistico <- glm(default ~ balance, data = datos, family = "binomial")

summary(modelo_logistico)
```

```
##
## Call:
## glm(formula = default ~ balance, family = "binomial", data = datos)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2697  -0.1465  -0.0589  -0.0221   3.7589
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
## balance      5.499e-03  2.204e-04   24.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1596.5  on 9998  degrees of freedom
## AIC: 1600.5
##
## Number of Fisher Scoring iterations: 8
```

beta 0 → ## (Intercept) -1.065e+01 3.612e-01 -29.49 <2e-16 ***

beta 1 → ## balance 5.499e-03 2.204e-04 24.95 <2e-16 ***

H0: beta0 = 0 (no significativo)
H1: beta0 !=0 (significativo)
p-valor < 0.05 entonces rechazo H0, beta0 es significativo
H0: beta1 = 0 (no significativo)
H1: beta1 !=0 (significativo)
pvalor < 0.05 entonces rechazo H0, beta1 es significativo

“

```
names(modelo_logistico)
```

```
##  [1] "coefficients"      "residuals"         "fitted.values"
##  [4] "effects"           "R"                 "rank"
##  [7] "qr"                "family"            "linear.predictors"
## [10] "deviance"          "aic"               "null.deviance"
## [13] "iter"              "weights"           "prior.weights"
## [16] "df.residual"       "df.null"           "y"
## [19] "converged"         "boundary"          "model"
## [22] "call"              "formula"           "terms"
## [25] "data"              "offset"            "control"
## [28] "method"            "contrasts"         "xlevels"
```

```
vcov(modelo_logistico)
```

```
##               (Intercept)       balance
## (Intercept)  1.304346e-01 -7.817111e-05
## balance     -7.817111e-05  4.856301e-08
```

```
# Ajuste del modelo
pchisq(deviance(modelo_logistico),df.residual(modelo_logistico) ,lower=FALSE)
```

```
## [1] 1
```

H0: Modelo actual (buen ajuste)

H1: Modelo saturado

Si p-valor < 0.05 rechazo Hipótesis nula

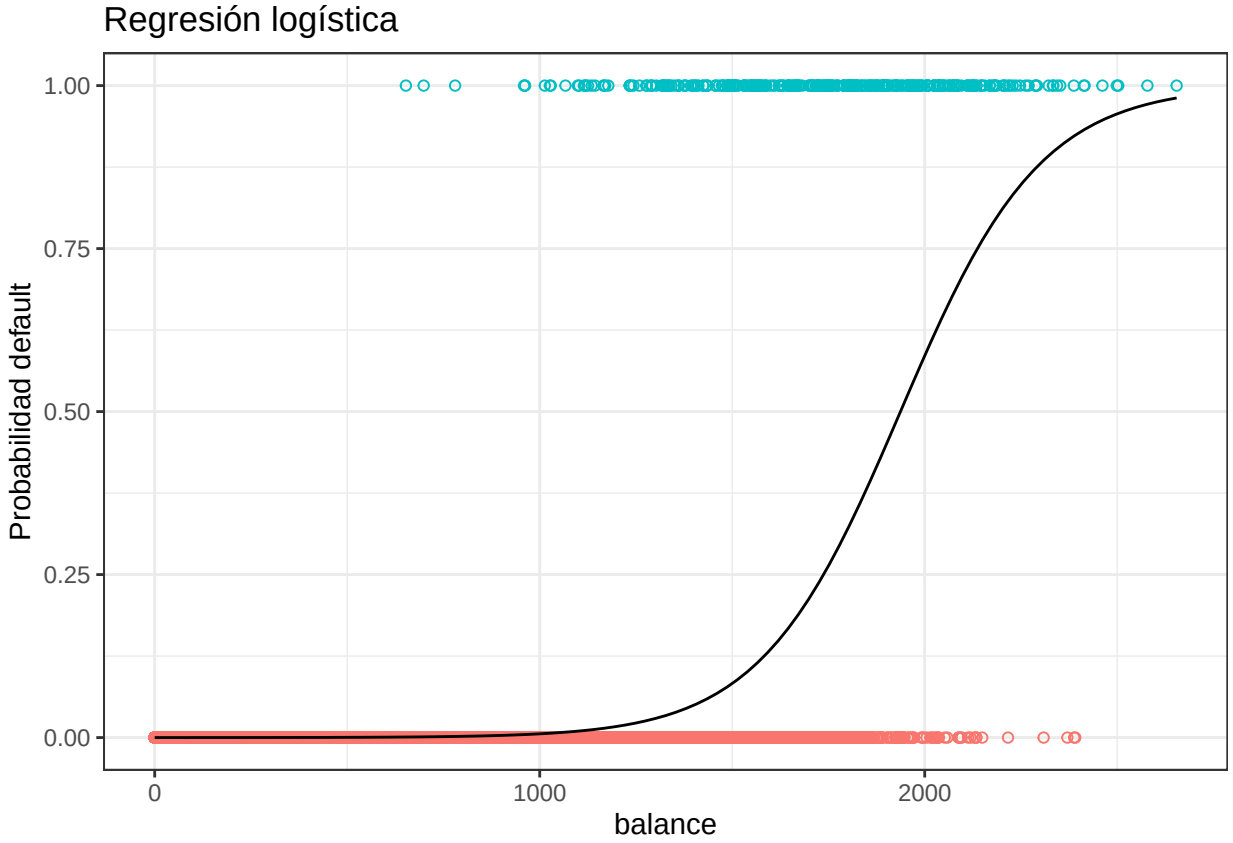
COmo p-valor $=1 > 0.05$, no rechazo la hipótesis nula, por lo tanto el modelo actual presenta buen ajuste.

```
# "Estimación"" de y para los valores observados:
ypred <- predict(modelo_logistico, type = "response")
```

```r
ypred[1:10]
```

```
##            1            2            3            4            5
## 1.305680e-03 2.112595e-03 8.594741e-03 4.344368e-04 1.776957e-03
##            6            7            8            9           10
## 3.704153e-03 2.211431e-03 2.016174e-03 1.383298e-02 2.366877e-05
```

```r
# Representación gráfica del modelo.
ggplot(data = datos, aes(x = balance, y = default)) +
  geom_point(aes(color = as.factor(default)), shape = 1) +
  stat_function(fun = function(x){predict(modelo_logistico,
                                          newdata = data.frame(balance = x),
                                          type = "response")}) +
  theme_bw() +
  labs(title = "Regresión logística",
       y = "Probabilidad default") +
  theme(legend.position = "none")
```



Sea $p_i = p(Y = 1|X = x_i)$, se define el ODD como:

$$ODD = p_i/(1 - p_i)$$

Suponga que la probabilidad de éxito es de 0.8, por lo que la probabilidad de fracaso es de 1 - 0.8 = 0.2. *Los ODDs (o razón de probabilidad) de éxitos se definen como el ratio entre la probabilidad de éxito y la probabilidad de fracaso p/q.

$$ODDS(exito) = probabilidad(éxito)/probabilidad(fracaso)$$

En este caso los ODDs de éxito son 0.8 / 0.2 = 4, lo que equivale a decir que se esperan 4 éxitos por cada fracaso. Nos define quién es más probable el éxito o el fracaso, según sus probabilidades respectivas.

La trasformación de probabilidades a ODDs es monótona, si la probabilidad aumenta también lo hacen los ODDs, y viceversa. El rango de valores que pueden tomar los ODDs es de [0,∞].

Dado que el valor de una probabilidad está acotado entre [0,1] se recurre a una trasformación logit (existen otras) que consiste en el logaritmo natural de los ODDs. Esto permite convertir el rango de probabilidad previamente limitado a [0,1] a $[-\infty, +\infty]$.

De esta forma,

$$LOG(ODDS) = ln(\frac{p(Y=1|x_i)}{1-p(Y=1|x_i)}) = \beta_0 + \beta_1 x_i = \eta,$$

en la regresión logística, tal como se ha descrito en la sección anterior, se modela la probabilidad de que la variable respuesta Y pertenezca al nivel de referencia 1 en función del valor que adquieran los predictores, mediante el uso de LOG of ODDs.

```
library(knitr)
p<- c(0.001,0.01, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.999,0.9999)
odds <- p/(1-p)
df <- data.frame(p ,odds , log(odds))

kable(df)
```

p	odds	log.odds.
0.0010	0.0010010	-6.9067548
0.0100	0.0101010	-4.5951199
0.2000	0.2500000	-1.3862944
0.3000	0.4285714	-0.8472979
0.4000	0.6666667	-0.4054651
0.5000	1.0000000	0.0000000
0.6000	1.5000000	0.4054651
0.7000	2.3333333	0.8472979
0.8000	4.0000000	1.3862944
0.9000	9.0000000	2.1972246
0.9990	999.0000000	6.9067548
0.9999	9999.0000000	9.2102404

Los ODDs y el logaritmo de ODDs cumplen que:

Si p(exito) = p(fracaso), entonces odds(exito) = 1

Si p(exito) < p(fracaso), entonces odds(exito) < 1

Si p(exito) > p(fracaso), entonces odds(exito) > 1

A diferencia de la probabilidad que no puede exceder el 1, los ODDs no tienen límite superior.

Si odds(exito) = 1, entonces logit(p) = 0

Si odds(exito) < 1, entonces logit(p) < 0

Si odds(exito) > 1, entonces logit(p) > 0

La transformación logit no existe para p = 0.

La razon de ODDS = $\frac{\frac{p(Y=1|X=x+1)}{1-p(Y=1|X=x+1)}}{\frac{p(Y=1|X=x)}{1-p(Y=1|X=x)}} = e^{\beta_1}$ indica el cambio en el logaritmo de ODDs de éxito debido al incremento de una unidad de X, es decir e^{β_1} es el cambio en el ODDS de éxito cuando el valor de la variable explicativa aumenta en una unidad.

En la literatura se sugiere interpretar el cociente de ODSS como el "aumento estimado" en la probabilidad de éxito asociado con un cambio unitario en el valor de la variable explicativa.

En general, el aumento estimado del cociente de ODDS, asociado con un cambio de d unidades en la variable predictora, es $e^{(d\beta^1)}$.

```
modelo_logistico$coefficients
```

```
##   (Intercept)       balance
## -10.651330614   0.005498917
```

El ODDS ((p/(1-p))) de un cliente realizar un fraude cuando tiene un balance de x_i+1 es $e^{0.005498917} = 1.005514$ veces el ODDS de un cliente realizar un fraude cuando tiene un balance x_i. De forma equivalente, el ODDS de un cliente realizar un fraude se incrementa en 0.05% por cada u.m. adicional en su balance.

En la práctica "se asume" que por cada u.m adicional en el balance aumenta (1- 1.005514)% =0.05%, la probabilidad de un cliente realizar un fraude.

Entonces por cada 10 u.m. adicionales en el balance de un cliente, $e^{0.005498917*10} = 1.056529$, el ODDS del cliente realizar un fraude se incrementa en 5%.

Por cada 100 u.m. adicionales en el balance de un cliente, $e^{(100*0.005498917)} = 1.733065$, entonces el ODDS de un cliente realizar un fraude se incrementa en 73%.

```
op <- par(mfrow = c(2, 2))
plot(modelo_logistico)
```