Master en Big Data. Fundamentos matemáticos del análisis de datos.

Sesión 8. Modelos lineales generalizados. Regresión Logística.

Fernando San Segundo

Curso 2020-21. Última actualización: 2021-09-30

COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI
DEPARTAMENTO DE MATEMÁTICA APLICADA

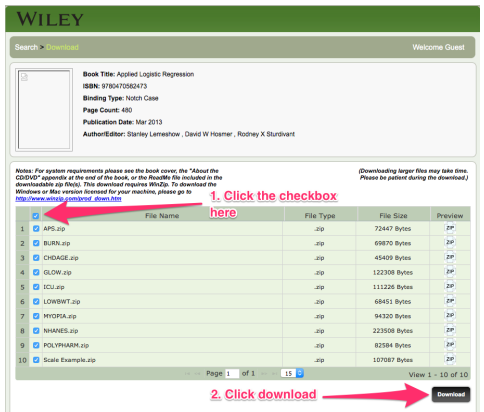
1 Modelos lineales generalizados (glm). Regresión Logística.

Sección 1

Modelos lineales generalizados (glm). Regresión Logística.

Relaciones del tipo $F \sim C$ para factores binarios.

- Vamos a estudiar ahora un modelo adecuado para relaciones F C, en las que una variable continua X es el predictor de una respuesta Y de tipo factor, que inicialmente suponemos *binario* (con dos niveles).
- Usaremos de ejemplo un conjunto de datos disponible en el sitio web de la editorial Wiley y procedente del libro *Applied Logistic Regression* de S. Lemeshow (Hosmer Jr et al. 2013). Introduce en el campo de búsqueda adecuado el ISBN del libro que es: 9780470582473 haz clic en *Search* y después haz clic en el enlace con ese número que aparecerá . A continuación marca la casilla para seleccionar todos los ficheros y después haz clic en *Download* como indica la figura. Descargarás entonces un fichero zip con todos los ficheros de datos necesarios. Nosotros vamos a usar uno de los ficheros que contiene, llamado `CHDAGE.txt` con datos sobre la existencia de enfermedad coronaria en un grupo de pacientes. Asegúrate de colocar ese fichero en la subcarpeta `datos` de tu directorio de trabajo.



Descripción del problema.

- Leemos los datos a un tibble. **Ejercicio:** ¿Por qué hemos usado `read_delim`?

```
library(tidyverse)
CHDdata <-  read_delim("./data/CHDAGE.txt", delim = "\t")
CHDdata %>% slice_head(n = 6)
```

```
## # A tibble: 6 x 3
##      ID   AGE   CHD
##   <dbl> <dbl> <dbl>
## 1     1    20     0
## 2     2    23     0
## 3     3    24     0
## 4     5    25     1
## 5     4    25     0
## 6     7    26     0
```

Hay tres variables: `ID` es simplemente un identificador y no la usaremos; `AGE` es la edad con valores enteros y `CHD` (de *coronary heart disease*) es un factor codificado como una variable binario que toma los valores 0 o 1 para indicar, respectivamente, la ausencia o presencia de enfermedad coronaria.

- Empezamos explorando los datos con `summary` (comprobamos que no hay datos ausentes):

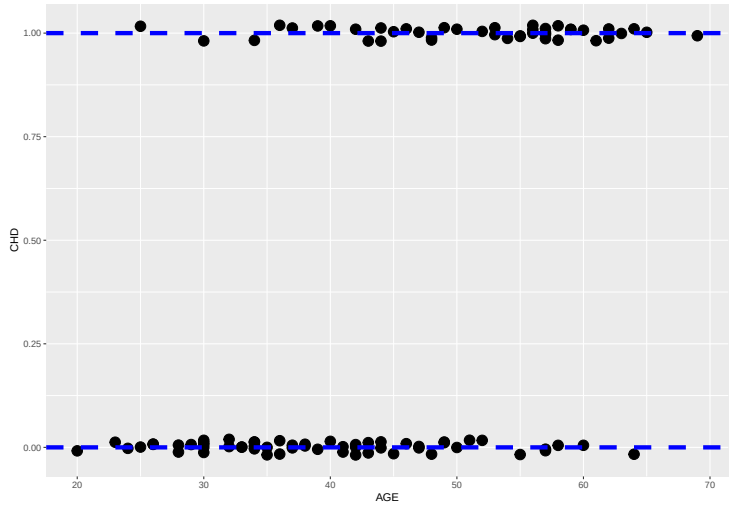
```
summary(CHDdata)
```

```
##        ID              AGE             CHD        
##  Min.   :  1.00   Min.   :20.00   Min.   :0.00  
##  1st Qu.: 25.75   1st Qu.:34.75   1st Qu.:0.00  
##  Median : 50.50   Median :44.00   Median :0.00  
##  Mean   : 50.50   Mean   :44.38   Mean   :0.43  
##  3rd Qu.: 75.25   3rd Qu.:55.00   3rd Qu.:1.00  
##  Max.   :100.00   Max.   :69.00   Max.   :1.00
```

Representando los datos.

- Para entender lo que queremos hacer vamos a usar una representación gráfica similar al diagrama de dispersión de la regresión lineal.

```
ggplot(CHDdata) +
  geom_point(aes(x = AGE, y = CHD, size=2),
             show.legend=FALSE, position = position_jitter(w = 0, h = 0.02)) +
  geom_hline(yintercept = 0:1, linetype = "dashed", color = "blue", size = 2)
```



Aunque Y es binario hemos *"agitado"* con jitter los puntos verticalmente para mejorar la visualización. Fíjate en que a medida que $X = AGE$ aumenta hay más valores de $Y = CHD$ iguales a 1. Como cabría esperar, la presencia de enfermedades coronarias aumenta con la edad. Esa es la *tendencia o señal* que estamos tratando de detectar o cuantificar expresándola a través de algún tipo de modelo.

Construcción del modelo.

- En una situación como esta no podemos usar un modelo lineal de regresión del tipo

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

  porque la respuesta Y no es continua, *solo toma dos valores*.
- La idea astuta que nos va a permitir avanzar en este caso es una que ya nos hemos encontrado antes. Cuando tratamos con una variable continua podemos agruparla en clases (en R con cut) para tratarla como un factor ordenado. Definimos unas franjas de edad con intervalos de 5 en 5 años, salvo el primero y el último porque tenemos pocos datos en los extremos. Usamos estos valores para cortar las edades y añadimos esa información a los datos:

```
AGEbreaks = c(20, seq(from = 30, to = 60, by = 5), 70)
CHDdata <- CHDdata %>%
  mutate(AgeGroup = cut(AGE, breaks = AGEbreaks, right = FALSE))
```

Analizando los datos agrupados...

- Hagamos una tabla de contingencia de CHD frente al grupo de edad:

```
(tabla1 = as.matrix(table(CHDdata$CHD, CHDdata$AgeGroup)))
```

```
##
##     [20,30) [30,35) [35,40) [40,45) [45,50) [50,55) [55,60) [60,70)
##   0       9      13       9      10       7       3       4       2
##   1       1       2       3       5       6       5      13       8
```

Ahí está de nuevo, visible, la señal. En la segunda fila de la tabla los valores aumentan claramente de izquierda a derecha (y en la primera ocurre al revés).

- Vamos a calcular la suma por columnas de esta tabla (es una tabla de frecuencias):

```
(sumaColumnas = colSums(tabla1))
```

```
## [20,30) [30,35) [35,40) [40,45) [45,50) [50,55) [55,60) [60,70)
##      10      15      12      15      13       8      17      10
```

Y dividimos la segunda fila de la `tabla1` anterior por estas sumas (redondeada a dos cifras). :

```
(probs = signif(tabla1[2, ] / sumaColumnas, 2))
```

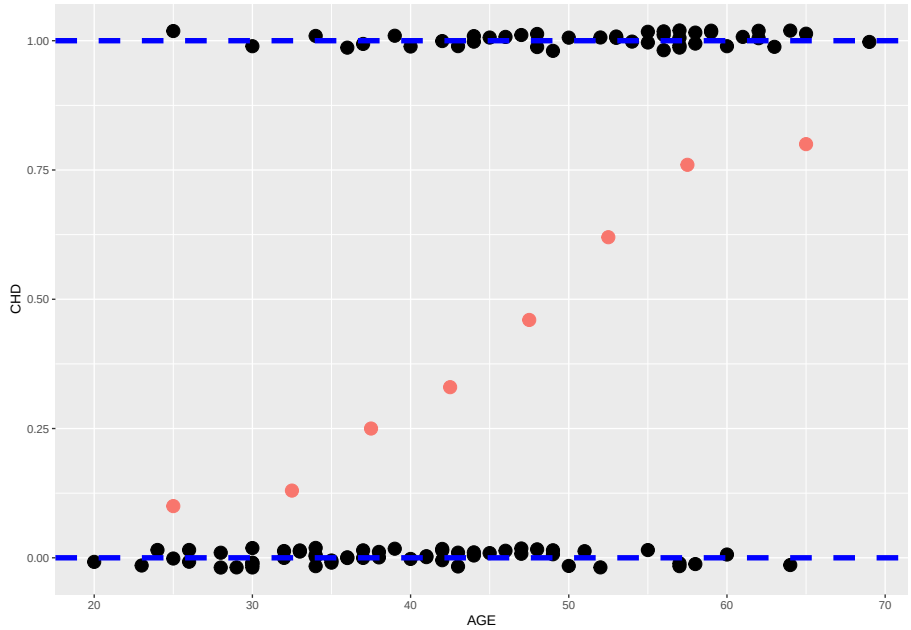
```
## [20,30) [30,35) [35,40) [40,45) [45,50) [50,55) [55,60) [60,70)
##    0.10    0.13    0.25    0.33    0.46    0.62    0.76    0.80
```

. . . para pensar en términos de probabilidades.

- Esta tabla

```
## [20,30) [30,35) [35,40) [40,45) [45,50) [50,55) [55,60) [60,70)
##    0.10    0.13    0.25    0.33    0.46    0.62    0.76    0.80
```

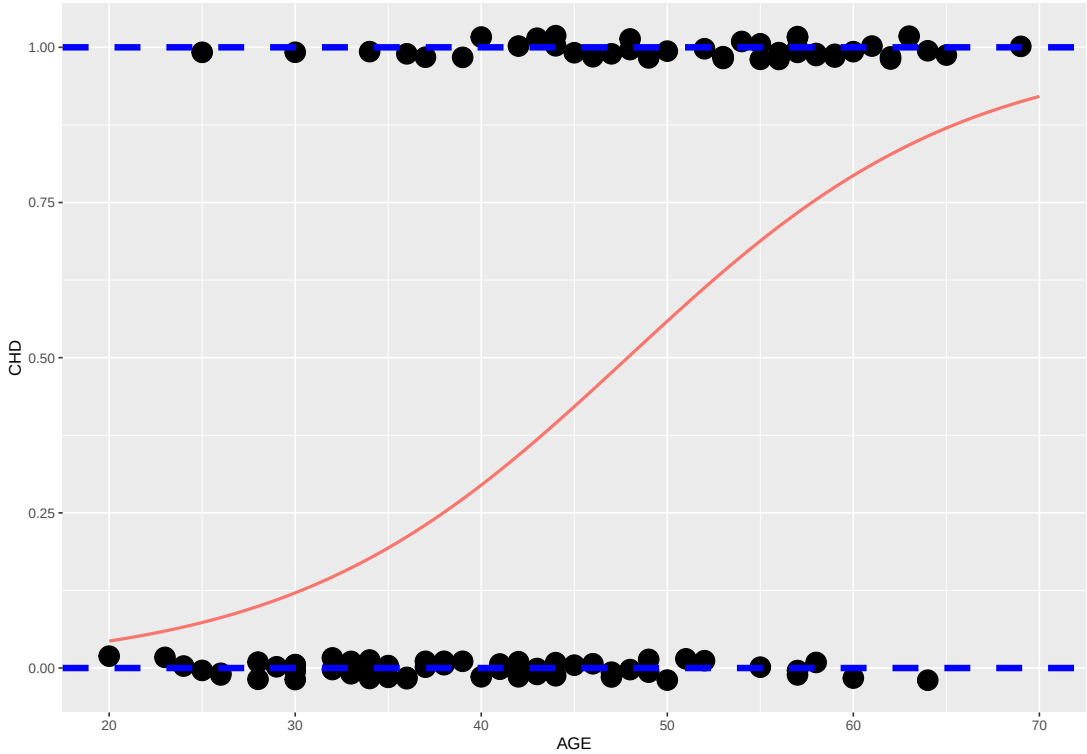
permite pensar en los datos en términos de probabilidades. Por ejemplo, el porcentaje de pacientes de 45 a 50 años con enfermedad coronaria es el 46 % y asciende al 76 % de 55 a 60 años. ¡Esa es la idea clave! Vamos a añadir esto al gráfico (ver código de la sesión):



Los puntos rojos indican la probabilidad de $Y = 1$ para cada intervalo de edades.

El modelo es una curva con forma de s.

- Los puntos rojos de la gráfica previa insinuan una curva con forma de s (*curva sigmoidal*) de izquierda a derecha, como esta (puedes ver el código para ver como la hemos dibujado, pero solo se entenderá después de aprender un poco más):



Esa curva representa el modelo que buscábamos para este tipo de situaciones. Es muy importante recordar que la coordenada vertical de los puntos de esa curva son **probabilidades condicionadas**. Es decir, un punto (x_0, p_{x_0}) de esa curva representa:

$$p_0 = P(Y = 1|X = x_0)$$

Curvas logísticas. Ajuste mediante la verosimilitud.

- Una vez entendido esto, el siguiente paso es más técnico. La familia de curvas sigmoidales que vamos a usar (puedes pensar en ellas como un sustituto de las rectas de regresión) es esta:

Curvas logísticas.

Dados dos números cualesquiera β_0, β_1 la curva logística correspondiente es:

$$f(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}.$$

Puedes familiarizarte con las propiedades de esta familia de curvas en este enlace.

- Se trata de elegir los valores de β_0 y β_1 que producen *la mejor curva logística posible* para ajustarla a nuestros datos. Esto recuerda a lo que hicimos en la regresión lineal. Allí usamos el método de mínimos cuadrados, pero ahora no podemos hacer esto, por razones técnicas (la estructura de error del problema no sigue una distribución normal, sino una binomial). El método general se basa en la función **verosimilitud (likelihood)**. En sentido amplio la verosimilitud de un modelo con parámetros (como β_0, β_1) se define como:

$$\mathcal{L}(\text{modelo con parámetros}) = P(\text{datos} \,|\, \text{dados los valores de los parámetros})$$

y para hallar β_0 y β_1 elegimos los que minimizan $-log(\mathcal{L})$, denominado *loglikelihood*.

Ajustando un modelo logístico con R. Función `glm`.

- La función verosimilitud es mucho más complicada que el error cuadrático medio que usamos en la regresión y no vamos a ver fórmulas para estimar β_0 y β_1. Dejaremos que R se encargue de estimarlos por nosotros. Usamos la función `glm` de *generalized linear models* (volveremos después sobre este nombre) En nuestro ejemplo:

```
(glmCHD = glm(CHD ~ AGE, family = binomial, data = CHDdata))
```

```
##
## Call:  glm(formula = CHD ~ AGE, family = binomial, data = CHDdata)
##
## Coefficients:
## (Intercept)          AGE
##     -5.3095       0.1109
##
## Degrees of Freedom: 99 Total (i.e. Null);  98 Residual
## Null Deviance:       136.7
## Residual Deviance: 107.4     AIC: 111.4
```

La llamada a `glm` y su respuesta son similares a las de `lm`, salvo por el argumento `family = binomial` que indica a `glm` que queremos un modelo logístico. En particular obtenemos *estimaciones* $\hat{\beta}_0$ y $\hat{\beta}_1$ de los parámetros del modelo logístico:

```
coefficients(glmCHD)
```

```
## (Intercept)         AGE
##  -5.3094534   0.1109211
```

Ahora es un buen momento para volver a examinar el código que genera esta figura y tratar de entenderlo.

Usando el modelo logístico para predecir.

- Con un modelo logístico y `predict` podemos hacer predicciones como hacíamos con un modelo lineal. Por ejemplo, ¿cuál es la probabilidad de padecer una enfermedad coronaria (probabilidad de $Y = 1$) que predice el modelo para un paciente de $X = 32$ años?

```
edadPredecir = data.frame(AGE = 32)
(probCHD = predict(glmCHD, newdata = edadPredecir, type = 'response'))
```

```
##         1
## 0.1467932
```

Es decir, un 15 %.
El argumento `type = 'response'` que hemos incluido en la llamada a `predict` es el que permite obtener una probabilidad. Si hubiéramos usado `type = 'link'` habríamos obtenido como respuesta el valor

$$\hat{\beta}_0 + \hat{\beta}_1 x_0$$

donde, en este ejemplo, $x_0 = 32$. Compruébalo comparando estos valores:

```
predict(glmCHD, newdata = edadPredecir, type = 'link')
coefficients(glmCHD)[1] + coefficients(glmCHD)[2] * 32
```

El valor $w = \hat{\beta}_0 + \hat{\beta}_1 x_0$ se denomina *log-odds* de x_0. Veamos qué son los *odds*

Odds y log-odds en el contexto de la regresión logística.

- Es una forma de entender la probabilidad común en el mundo anglosajón (especialmente en el contexto de las apuestas). La Regla de Laplace dice:

$$P(A) = \frac{\text{núm. de sucesos elementales favorables a } A}{\text{núm. total de sucesos elementales}}.$$

  En la misma situación los *odds a favor* de A (no hay una traducción establecida, a mi me gusta *posibilidades*) se calculan como:

$$O_A = \frac{\text{núm. de sucesos elementales favorables a } A}{\text{núm. de sucesos elementales } \textbf{contrarios} \text{ a } A}.$$

  Por ejemplo una probabilidad de $\frac{3}{11}$ se convierte fácilmente en unos odds de $\frac{3}{11-3} = \frac{3}{8}$ (apuestas 3 a 8). Los odds correspondientes a una probabilidad p son:

$$O_p = \frac{p}{1-p}$$

- ¿Como usamos esto en la regresión logística? El modelo se puede escribir

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

  y si llamamos $w = \beta_0 + \beta_1 x$ y despejamos se obtienen los **log-odds** de x:

$$w = \ln\left(\frac{p}{1-p}\right) = \ln\left(O_p\right)$$

Interpretación del coeficiente β_1 del modelo de regresión logística.

- Ya podemos interpretar β_1: al aumentar x en una unidad los log-odds $\beta_0 + \beta_1 x$ aumentan en β_1. Después podemos traducir los odds en términos de probabilidades, como hemos visto.
- Por eso `predict` permite obtener el resultado como *log-odds*. A veces eso es preferible porque la relación de los log-odds con X es más simple que con las probabilidades. Veamos lo que pasa al considerar valores consecutivos de la edad y obtener las predicciones de log-odds y probabilidades:

```
edades = data.frame(AGE = 50:55)
probabilidades = predict(glmCHD, newdata = edades, type = 'response')
diff(probabilidades)
```

```
##           2           3           4           5           6
## 0.02714072 0.02662822 0.02597167 0.02518590 0.02428792
```

```
logOdds = predict(glmCHD, newdata = edades, type = 'link')
diff(logOdds)
```

```
##         2         3         4         5         6
## 0.1109211 0.1109211 0.1109211 0.1109211 0.1109211
```

```
coefficients(glmCHD)[2]
```

```
##       AGE
## 0.1109211
```

Como se ve, al aumentar la edad en un año el incremento de las probabilidades (calculado con `diff`) no es constante, pero el de los log-odds sí y coincide con el coeficiente estimado $\hat{\beta}_1$.

La idea detrás del modelo lineal generalizado (glm).

- La idea clave al construir el modelo de regresión logística es pasar de pensar en valores de Y a pensar en probabilidades $p = P(Y = 1|X = x)$; *el modelo predice probabilidades*. Podemos dar otro paso con una perspectiva más general.
- Recordemos que la *variable respuesta condicionada* $(Y|X = x)$ es una variable binaria con valores 1 (probabilidad p) y 0 (con probabilidad q). Es una variable de Bernouilli y su media es precisamente $\mu_{Y|X=x} = p$. Por tanto podemos pensar que en realidad *el modelo predice medias de* Y (una para cada valor de X).
- Recordando la expresión del modelo de regresión lineal:

$$Y = \beta_0 + \beta_1 X + \epsilon, \qquad \text{con} \quad \epsilon \sim N(0, \sigma)$$

la media de la variable respuesta condicionada $(Y|X = x)$ (usando que $\mu_\epsilon = 0$) es:

$$\mu_{Y|X=x} = \beta_0 + \beta_1 x = \hat{Y}(x)$$

Otra vez: *el valor que el modelo predice es* $\mu_{Y|X=x}$*, la media de* Y *para ese valor de* X.

Formalizando el glm.

- Podemos convertr esa idea en una definición. Un **modelo lineal generalizado** consta de tres componentes:
  1. Una **función de distribución** de probabilidad f con una distribución conocida para modelizar la variable respuesta condicionada. Técnicamente, la distribución f debe ser de la *familia exponencial*, una familia muy amplia de variables aleatorias que incluye la normal, la binomial, etc.
  2. Un **predictor lineal** $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$ que depende de las variables predictoras $X_1, \ldots, X_p$ (pueden ser continuas o factores usando variables índice).
  3. Una **función de enlace (link function)** g que sirve para lo que esperamos del modelo: *predecir medias de* Y (una media para cada valor de X).

Ejemplos de glm

- *Ejemplo 1:* en el modelo de regresión lineal simple la distribución f que usamos es la de normal $N(0, \sigma)$, el predictor lineal es $\beta_0 + \beta_1 X$ mientras que la función de enlace es la identidad $g(\mu) = \mu$. Por eso este modelo es el más simple, por la sencillez del enlace.
- *Ejemplo 2:* en el modelo de regresión logística con una variable la distribución f es la Bernouilli con probabilidad $p = \mu_{Y|X=x}$ que hemos visto, el predictor lineal vuelve a ser $\beta_0 + \beta_1 X_1$ mientras que la función de enlace es $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$, la función log-odds que es la inversa de la transformación logística.
- Se pueden definir otros modelos lineales generalizados cambiando estas componentes para por ejemplo modelizar una relación $Y \sim X$ donde Y es una variable de tipo Poisson (regression de Poisson) o para modelizar una variable respuesta de tipo factor politómico con más de dos niveles (regresión multinomial). Ver Wikipedia y Regression Models de B.Caffo.

Referencias para la sesión

Enlaces

- Código de esta sesión
- Regression Models for Data Science in R, de Brian Caffo.

Bibliografía

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.