Práctica 5

Maximiliano Vaca Montejano
364897
maximiliano.vaca@uabc.edu.mx

Profesor: Andrés García Medina
andres.garcia.medina@uabc.edu.mx
Fecha de entrega: jueves 9 de mayo, 2024 (12pm).

Ejercicio 1

Los siguientes datos muestran las edades (en años) e indicadores de presencia o ausencia de daño significativo en la coronaria de 100 individuos seleccionados para participar en el estudio.

Deseamos establecer una relacion entre la edad de una persona y su propension a padecer un problema en la coronaria

(a) Encuentre el valor esperado de las betas asociadas por medio de optimización
directa (IRLWLS) y compare sus resultados con ayuda de la función glm

In [18]:
```
data <- read.csv('datos/coronaria.csv')
colnames(data) <- c("edad", "CHD")

head(data, 10)
```

A data.frame: 10 × 2

	edad<int>	CHD<int>
1	20	0
2	23	0
3	24	0
4	25	0
5	25	1
6	26	0
7	26	0
8	28	0
9	28	0
10	29	0

In [19]:
```
#condiciones
b0 = c(-10, 0.2); b0 = t(t(b0))
tol = 1e-6
n = 100

#definimos X y Y
Y = data$CHD
X = cbind(1, data$edad) # matriz de diseño

#print("X:"); head(X, 5)

#X transpuesta
Xt = t(X)

#print("dim X:"); dim(X)
#print("dim b0:"); dim(b0)

# u
u = 1 / (1 + exp(- X%*%b0))

#print("u:"); dim(u); head(u, 5)
```

definimos W

In [20]:
```
v = u*(1-u)

W = diag(as.vector(v))

dim(W)
```

100 · 100

para b1

In [21]:
```
# b1 = b0 + solve(t(X)%*%W%*%X) %*% t(X) %*%(Y-u)

# b1
```

para cada b

In [22]:
```
i = 0

#b <- c(b0) # no

b <- list(b0)
#as.vector(b)

while ((m > tol) & (i < 100)) {
    i <- i + 1

    b1 <- b0 + solve(t(X)%*%W%*%X) %*% t(X) %*%(Y-u)

    m <- sqrt(sum((b1 - b0)^2))
    b0 <- b1

    #nueva u
    u = 1 / (1 + exp(- X%*%b0))
    #nueva W
    v = u*(1-u)
    W = diag(as.vector(v))

    b <- c(b, list(b0))
}

#b
b <- as.matrix(b)
```

In [23]:
```
b0
```

A matrix: 2 × 1 of type dbl

-5.3094534
0.1109211

In [24]:
```
head(b, 5)
```

A matrix: 5 × 1

-10.0, 0.2
-1.49720569, 0.03767488
-4.38035829, 0.09253679
-5.2216853, 0.1091822
-5.3085973, 0.1109042

excluimos el b0

In [25]:
```
b <- as.matrix(b[2:dim(b)[1]])

head(b, 5)
tail(b, 5)
```

A matrix: 5 × 1

-1.49720569, 0.03767488
-4.38035829, 0.09253679
-5.2216853, 0.1091822
-5.3085973, 0.1109042
-5.3094533, 0.1109211

A matrix: 5 × 1

[2,]	-4.38035829, 0.09253679
[3,]	-5.2216853, 0.1091822
[4,]	-5.3085973, 0.1109042
[5,]	-5.3094533, 0.1109211
[6,]	-5.3094534, 0.1109211

comparamos con glm

In [26]:
```
#comparar con glm

glm_model <- glm(CHD ~ edad, data = data, family = binomial)

b0_glm <- glm_model$coefficients
b0_glm
```

(Intercept): -5.30945337391173 **edad:** 0.110921142206757

In [27]:
```
abs(b0[1] - b0_glm[1])
abs(b0[2] - b0_glm[2])
```

(Intercept): 7.31681382148963e-12
edad: 1.43524081508417e-13

guardamos los betas obtenidos para mas tarde, en b_0 y b_1

In [28]:
```
b_0 = b0[1] ; b_1 = b0[2]
```

difieren por muy poco, parece evidente que nuestra funcion de recurrencia converge al resultado del modelo

(b) Encuentre la desviación estándar de las betas asociadas por medio de optimización directa (IRLWLS) y compare sus resultados con ayuda de la función glm. (20 pts)

In [29]:
```
dim(t(X)); dim(W); dim(X)
```

2 · 100

100 · 100

100 · 2

In [30]:
```
V = solve(t(X)%*%W%*%X)
print('V: '); V
print('diag(V)'); diag(V)
print('sqrt(diag(V))'); sqrt( diag( V ) )

print('glm model:')
summary(glm_model)$coefficients
```

[1] "V: "
A matrix: 2 × 2 of type dbl

1.28517284	-0.0266779195
-0.02667702	0.0005788757

[1] "diag(V)"
1.28517285557272 -0.000578875702331837

[1] "sqrt(diag(V))"
1.13365463681525 0.0240059835874979

[1] "glm model:"
A matrix: 2 × 4 of type dbl

	Est.Imate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.3094534	1.1336365	-4.683488	2.823338e-06
edad	0.1109211	0.0240598 2	4.610224	4.022356e-06

son muy similares

de clase

$$\frac{momios\ de\ y=1\ con\ Z_i+1}{momios\ de\ y=1\ con\ Z_i}=e^{\beta_1}$$

usamos nuestro β_1
si comparamos dos individuos con 16 años de diferencia:

$$\frac{momios\ de\ y=1\ con\ Z_i+16}{momios\ de\ y=1\ con\ Z_i}=e^{16*\beta_1}$$

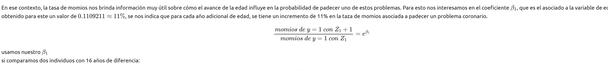
In [31]:
```
exp(16 * b_1)
```

5.89873708913651

(c) Grafique los datos en conjunto con la solución del modelo de regresión lineal y un modelo lineal. Se espera que se obtenga una gráfica como la mostrada en la figura 1(a) (20 pts)

In [32]:
```
#graficar los datos, la curva ajustada y regresion lineal
plot(data$edad, data$CHD, col = "blue", pch = 19, xlab = "Edad", ylab = "CHD")
curve(1 / (1 + exp(-b_0 - b_1*x)), add = TRUE, col = "red")
abline(glm_model, col = "green")
```

(d) Utilice la librería ROCR para graficar la curva ROC y determinar el area bajo la curva (AUC). Se espera que obtenga un grafico como el de la figura 1(b) (20 pts)



In [33]:
```
#u
```

In [34]:
```
library(pROC)

# calculamos la curva ROC para las predicciones del modelo u="miu" con respecto
# a las verdaderas etiquetas data$CHD.
u <- as.vector(u)
roc_obj <- roc(data$CHD, u)

# se grafica la curva ROC.
plot(roc_obj, col = "blue", main = "ROC")

# calculmos el AUC de la curva ROC.
# Con esto sabr el modelo puede distinguir entre las diferentes clases, de 0 a 1
auc(roc_obj)
```

```
Setting levels: control = 0, case = 1

Setting direction: controls < cases
```

0.799877600979192

obtenemos un valor de AUC=0.799877600979192, el cual se considera aceptable

(e) Discuta las implicaciones de los resultados del caso de estudio particular a través de la tasa de momios y de lo obtenido en los incisos anteriores.
Nota: mínimo 200 palabras. (10 pts).

En esta práctica realizamos un análisis para conocer la relación entre la edad y la propensión a padecer un problema coronario, para esto aplicamos un modelo lineal generalizado de regresión logística.
Los resultados obtenidos para los coeficientes de regresión son:

$$\beta_0 = -5.3094534$$
$$\beta_1 = 0.1109211$$

donde β_0 es el intercepto y β_1 es el asociado a la edad.

Al realizar un análisis sobre las desviaciones estándar de los estimadores, nos encontramos con valores no muy grandes, lo cual verifica su confiabilidad, respaldada además con el hecho de haber obtenido un valor muy similar al aplicar el modelo desde la librería.

$$Std.\ Error(\beta_0) = 1.1336546366815250$$
$$Std.\ Error(\beta_1) = 0.0240598358749979$$

En ese contexto, la tasa de momios nos brinda información muy útil sobre cómo el avance de la edad influye en la probabilidad de padecer uno de estos problemas. Para esto nos interesamos en el coeficiente β_1, que es el asociado a la variable de edad. Al haber obtenido para este un valor de $0.1109211 \approx 11\%$, se nos indica que para cada año adicional de edad, se tiene un incremento de 11% en la taza de momios asociada a padecer un problema coronario.

$$\frac{momios\ de\ y=1\ con\ Z_i+1}{momios\ de\ y=1\ con\ Z_i}=e^{\beta_1}$$

usamos nuestro β_1
si comparamos dos individuos con 16 años de diferencia:

$$\frac{momios\ de\ y=1\ con\ Z_i+16}{momios\ de\ y=1\ con\ Z_i}=e^{16*\beta_1}=5.898737089136513$$

Por último, al graficar la curva ROC, lo que nos interesa es determinar su área bajo la curva, la cual nos habla sobre qué tan bueno es el modelo en discriminar entre las clases al realizar predicciones. Para esta última obtuvimos un valor de AUC=0.799877600979192, el cual se considera aceptable. Un valor alrededor de 0.5 habría indicado un rendimiento casi al azar, mientras que uno menor a esto, completamente deficiente.

In [ ]:
```

```