Biostatistics Exercises

www.nataliejulian.com

Regresión Logística

El archivo `datosUCI.txt` contiene información sobre un estudio de sobrevivencia de pacientes en la UCI (Unidad de Cuidados Intensivos). El objetivo principal del estudio es desarrollar un modelo de regresión logística para predecir la probabilidad de sobrevivencia de estos pacientes. Las variables consideradas en el estudio son las siguientes:

- ID (código de identificación)
- STA (estado vital, 1=muerto, 0=vivo)
- AGE (edad en años)
- SEX (sexo, 1=mujer, 0=hombre)
- RACE (raza, 1=blanco, 2=negro, 3=otra)
- CAN (problemas de cáncer, 2=si, 1=no)
- CRN (problemas de riñon, 2=si, 1=no)
- INF (probable infección al ingresar a la UCI, 2=si, 1=no)
- CPR (resucitación pulmonar CPR 2=si, 1=No)
- SYS (presión sanguínea sistólica al ingreso en mm Hg)
- HRA (presión del corazón al ingreso en pulsaciones/min)
- PRE (Admisión previa a la UCI dentro de 6 meses (2=si, 1=No)
- TYP (Tipo de admisión, 1=electiva, 2=emergencia)

entre otras.

1. En la base de datos, la variable respuesta de principal interés es el estado vital, STA. Médicos asociados con el estudio creen que un factor determinante de sobreviviencia es la edad de admisión del paciente, AGE.

a) Escriba la ecuación para el modelo de regresión logística de STA sobre AGE. Escriba la ecuación para la transformación logit de este modelo de regresión logística? ¿Qué característica de la variable respuesta STA, nos lleva a considerar un modelo de regresión logística y no un modelo de regresión lineal usual, para describir la relación entre STA y AGE?

La ecuación para la transformación logit del modelo de regresión logística de STA sobre AGE considerando intercepto y también considerando que STA=0 es el éxito (es decir, que la persona sobreviva) es la siguiente:

$$\frac{P(STA=0|AGE)}{1-P(STA=0|AGE)} = exp(\beta_0 + \beta_1 AGE)$$

Note que este modelo se construye a partir de la razón de chances del éxito el que en este caso, es sobrevivir (STA=0). Así, el modelo anterior puede representarse de manera equivalente como una función lineal del log-odds de la sobrevivencia dado valores de las covariables (AGE):

$$log\left(\frac{P(STA=0|AGE)}{1-P(STA=0|AGE)}\right) = \beta_0 + \beta_1 AGE$$

La ecuación para obtener $\hat{\beta}$ es la siguiente:

$$\beta^1 = \beta^0 + [X^T W X]^{-1} X^T (X - \mu)$$

Donde X corresponde a la matriz de diseño del modelo (bajo un modelo con intercepto la primera columna es constante y la segunda columna es la covariable AGE), W corresponde a la matriz diagonal de pesos $n_i\pi_i(1-\pi_i)$ donde $\pi_i = \frac{exp(\beta_0+\beta_1 AGE_i)}{1+exp(\beta_0+\beta_1 AGE_i)}$ y μ es un vector tal que $\mu_i = n_i\pi_i$ que nace de la naturaleza de la variable STA como una variable que indica éxito o fracaso, donde cada realización de esta variable es un ensayo Bernoulli

Es evidente que realizar un modelo de regresión lineal no posee sustento teórico alguno, dada la naturaleza de la variable STA, pues ésta al ser una variable dicotómica necesita ser modelada de manera tal de cuantificar la probabilidad de éxito o no éxito (lo que realiza una regresión logística) mientras que una regresión lineal descompone la media de la variable respuesta en una combinación lineal de las covariables, lo cual no corresponde a una probabilidad y por lo tanto, pudiera entregar valores que se escapan del intervalo [0,1]. Existiría una clara inconsistencia en los supuestos al utilizar una regresión lineal usual para modelar la variable STA.

b) Obtenga un gráfico de dispersión de STA versus AGE. Comente.

```
> plot(AGE,STA,pch="",
+      main="Indicatriz de sobrevivencia en UCI por Edad de admisión",
```

```
+      xlab="Edad de admisión",ylab="Indicatriz de sobrevivencia")
> points(AGE[STA==0],STA[STA==0],col="thistle3",pch=15)
> points(AGE[STA==1],STA[STA==1],col="dodgerblue4",pch=15)
> legend("left",legend=c("Sobrevive","No sobrevive"),
+        col=c("thistle3","dodgerblue4"),cex=0.8,pch=15)
>

> table(STA)

STA
  0   1
160  40

> summary(AGE)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.00   46.75   63.00   57.55   72.00   92.00
```

Mirando el gráfico es posible observar de manera inmediata, el por qué no tendría sentido utilizar una regresión lineal con la variable respuesta STA como la indicatriz de sobrevivencia de los pacientes, debido a la naturaleza dicotómica de la variable respuesta. Adicionalmente, es posible observar que los casos donde no hay sobrevivencia del paciente se concentran en edad de admisión desde los 40 años (excepto algunos pocos casos cercanos a los 20 años). Mientras que los pacientes que sobreviven se distribuyen a lo largo de todos los valores de edad de admisión (desde 16 hasta 92 años). También podemos observar que existen más casos donde los pacientes sobrevivieron, registrando 160 casos en los que sobrevive el paciente versus 40 donde se observó la no sobrevivencia del paciente.

c) Usando los intervalos [15,24], [25,34], [35,44], [45,54], [55,64], [65,74], [75,84], [85,94] para AGE, calcule el promedio de STA para los sujetos dentro de cada intervalo de edad. Grafique estos valores (es decir, las medias de STA) versus los puntos medios de los intervalos de edad usando el mismo conjunto de ejes que en 1 (b). Comente.

```
> datosUCI$intervalos<-cut(AGE,
+                          breaks = c(15,24,34,44,54,64,74,84,94),
+                          labels = paste("Intervalo", 1:8), right=TRUE)
> library(dplyr)
> df<-datosUCI %>%
+   group_by(intervalos) %>%
+   summarise(prop=mean(STA))
> df$puntosmedios<-c((15+24)/2, (25+34)/2, (35+44)/2, (45+54)/2,
+                    (55+64)/2, (65+74)/2, (75+84)/2, (85+94)/2)
> df
```

```
# A tibble: 8 x 3
  intervalos    prop puntosmedios
  <fct>        <dbl>        <dbl>
1 Intervalo 1 0.0769         19.5
2 Intervalo 2 0              29.5
3 Intervalo 3 0.182          39.5
4 Intervalo 4 0.2            49.5
5 Intervalo 5 0.205          59.5
6 Intervalo 6 0.18           69.5
7 Intervalo 7 0.3            79.5
8 Intervalo 8 0.455          89.5
```

```
> library(ggplot2)
> df%>%
+   ggplot(aes(x=puntosmedios, y=prop, label=intervalos)) +
+      geom_point() +
+      geom_segment(aes(xend=c(tail(puntosmedios, n=-1), NA),
+                       yend=c(tail(prop, n=-1), NA))) +
+   scale_y_continuous(limits=c(0,1))+
+   xlab("Puntos medios rangos etarios de admisión")+
+   ylab("Proporción de pacientes fallecidos")+
+   ggtitle("Proporción de pacientes fallecidos por rango etario")+
+   theme_light()+theme(plot.title=element_text(hjust=0.5, size=16))
```

El promedio de la variable STA dada la codificación binaria que posee (STA=1 muerte) equivale a la proporción de fallecidos, por ende, lo que se muestra en el gráfico es por intervalo de Edad de admisión la proporción de fallecidos. Observando el gráfico es posible ver que en generan se observaría una tendencia creciente de la proporción de pacientes fallecidos por edad de admisión, es decir, en general, existirían mayores probabilidades de fallecer del paciente a medida que entra en una categoría de admisión más alta. Sin embargo, note que existen algunas bajas en el intervalo [25,34] años y [65,74] años. También es posible observar que la máxima proporción de fallecidos por intervalo es de aproximadamente 0.5 es decir, la máxima mortalidad por grupo se observó en el intervalo [85,94] donde aproximadamente la mitad de los pacientes en ese rango etario de admisión registra la no sobrevivencia.

d) Escriba una expresión para la verosimilitud y la log-verosimilitud para el modelo de regresión logística de 1(a) usando los datos no agrupados (es decir considerando los $n = 200$ datos). Obtenga expresiones para las dos ecuaciones de verosimilitud.

Considerando que STA=0 es sobrevivir y lo que nos interesa es estudiar la **sobrevivencia** se definió el modelo de la siguiente forma:

$$log\left(\frac{P(STA=0|AGE)}{1-P(STA=0|AGE)}\right) = \beta_0 + \beta_1 AGE$$

Y $\pi_i = \frac{exp(\beta_0+\beta_1 AGE_i)}{1+exp(\beta_0+\beta_1 AGE_i)}$ la probabilidad de sobrevivencia estimada en función de AGE.

Defina como $\widehat{STA} = 1 - STA$ (esto pues nos interesa analizar la sobrevivencia). Así, la función de verosimilitud puede escribirse como sigue:

$$L\left(\beta_0, \beta_1 | STA, AGE\right) = \prod_{i=1}^{200} \pi_i^{\widehat{STA}_i} \left(1-\pi_i\right)^{1-\widehat{STA}_i}$$

Esta expresión surge de la naturaleza Bernoulli de la variable $\widehat{STA}$.

Y la función de log-verosimilitud $l\left(\beta_0, \beta_1 | STA, AGE\right)$ se puede expresar como sigue:

$$= \sum_{i=1}^{n} \widehat{STA}_i \log\left(\pi_i\right) + \left(1 - \widehat{STA}_i\right) \log\left(1-\pi_i\right)$$

$$= \sum_{i=1}^{n} \widehat{STA}_i \log\left(\frac{exp(\beta_0+\beta_1 AGE_i)}{1+exp(\beta_0+\beta_1 AGE_i)}\right) + \left(1 - \widehat{STA}_i\right) \log\left(1 - \frac{exp(\beta_0+\beta_1 AGE_i)}{1+exp(\beta_0+\beta_1 AGE_i)}\right)$$

$$= \sum_{i=1}^{n} \widehat{STA}_i \log\left(\frac{exp(\beta_0+\beta_1 AGE_i)}{1+exp(\beta_0+\beta_1 AGE_i)}\right) + \left(1 - \widehat{STA}_i\right) \log\left(\frac{1}{1+exp(\beta_0+\beta_1 AGE_i)}\right)$$

La maximización de la log-verosimilitud al no poseer una forma cerrada, sugiere que $\hat{\beta}$ se obtiene utilizando algoritmos de aproximación numérica, Fisher-Scoring, Newton Raphson u otros.

e) Usando una rutina de regresión logística, obtenga las estimaciones de máxima verosimilitud para los parámetros del modelo de regresión logística de 1(a). Estos estimadores deben estar basados en los $n = 200$ datos (es decir, no agrupados). Usando las estimaciones, escriba la ecuación para los valores ajustados, es decir, las probabilidades estimadas. Grafique la ecuación para los valores ajustados sobre los ejes utilizados en los gráficos obtenidos en 1(b) y 1(c).

```
> datosUCI$barSTA<-factor(1-STA)
> model<-glm(barSTA ~ AGE, family = binomial, data = datosUCI)
> #Por defecto al usar binomial se asume un link logit
>
> summary(model)  #interpretación de los coeficientes en términos de los odds

Call:
glm(formula = barSTA ~ AGE, family = binomial, data = datosUCI)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2854   0.3905   0.6145   0.7391   0.9536

Coefficients:
```

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.05851    0.69608   4.394 1.11e-05 ***
AGE         -0.02754    0.01056  -2.607  0.00913 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 200.16  on 199  degrees of freedom
Residual deviance: 192.31  on 198  degrees of freedom
AIC: 196.31

Number of Fisher Scoring iterations: 4

> coef(model)

(Intercept)         AGE
 3.05851323 -0.02754261
```

Para maximizar la log-verosimilitud se utilizó el algoritmo de Fisher Scoring, en el cual se convergió en la cuarta iteración. Se obtuvieron las siguientes estimaciones:

$$\hat{\beta}_0 = 3.05851 \qquad \hat{\beta}_1 = -0.02754$$

De lo cual, es posible obtener las probabilidades estimadas:

$$\hat{\pi}_i = \frac{exp(3.05851 - 0.02754 AGE_i)}{1 + exp(3.05851 - 0.02754 AGE_i)}$$

Cuando la edad aumenta en una unidad el logaritmo de las chances varía en -0.02754 (es decir, las chances disminuyen conforme aumenta la edad).

```
> plot(AGE[order(AGE)], model$fitted.values[order(AGE)], type="o",
+      xlab="Edad", ylab="Probabilidad estimada de sobrevivencia",
+      main="Probabilidad de sobrevivencia por Edad en UCI", pch=19,
+      col="lightsteelblue")
```

Complemento: Utilizar un enlace logit es adecuado cuando se observa simetría en 0.5, es decir, acercarse al valor 1 o al valor 0 tiene (o resulta razonable que tenga) la misma velocidad y por lo tanto se observa un punto de simetría en 0.5. Sin embargo, en este caso, no podemos aseverar que exista la misma rapidez de acercarse a la probabilidad 1 de sobrevivir que a la probabilidad 0 de sobrevivir; se tienen tan sólo 40 registros con pacientes que fallecieron, mientras que 160 sobrevivieron es decir, las categorías se encuentran desbalanceadas. El enlace log-log complementario se

sugiere cuando existe un evento que posee una mayor probabilidad considerable respecto al otro evento:

```
> model2<-glm(barSTA ~ AGE, family = binomial(link=cloglog), data = datosUCI)
> summary(model2)

Call:
glm(formula = barSTA ~ AGE, family = binomial(link = cloglog),
    data = datosUCI)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3296   0.3704   0.6267   0.7440   0.9179

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.244391   0.287650   4.326 1.52e-05 ***
AGE         -0.012955   0.004674  -2.772  0.00557 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 200.16  on 199  degrees of freedom
Residual deviance: 192.30  on 198  degrees of freedom
AIC: 196.3

Number of Fisher Scoring iterations: 5

> #Las interpretaciones de los coeficientes cambian de odds a hazard
>
>
> #La curva con link cloglog se observa levemente más suavizada
> plot(AGE[order(AGE)], model2$fitted.values[order(AGE)], type="o",
+      xlab="Edad", ylab="Probabilidad estimada de sobrevivencia",
+      main="Probabilidad de sobrevivencia por Edad en UCI", pch=19,
+      col="lightsteelblue")
> #En términos de devianza residual, no se observan mayores diferencias
> #entre ambos modelos
> deviance(model)

[1] 192.3064

> deviance(model2)
```

```
[1] 192.3
```

f) Resuma (describa en palabras) los resultados presentados en los gráficos de partes 1(b), 1(c) y 1(e).

En el gráfico 1(b) es posible ver la distribución de la variable dicotómica STA en la cual se observó una clara concentración de aquellos pacientes que fallecieron en la UCI en edades más altas, mientras que las edades de los pacientes que sobreviven se distribuyen en todo el rango de valores que toma la variable Edad de admisión. Luego en el gráfico 1(c) se observó la tasa de fallecimientos por rango etario de admisión, en el cual se observa la tendencia incremental de la tasa de fallecimientos en términos de rango etario de admisión. En el gráfico en 1(e), se modeló la probabilidad de sobrevivencia, este gráfico es complementario con el gráfico 1(c), mientras que en 1(c) se observa un aumento en la tasa de fallecimientos por rango etario, en el el gráfico 1(e) se observa una disminución de la probabilidad de sobrevivencia estimada por la edad de admisión.

g) Usando la salida del paquete de regresión logística utilizado en 1(e), evalúe la significancia del coeficiente asociado a AGE usando el test de razón de verosimilitud, el test de Wald y el test de Score. ¿Qué supuestos son necesarios para que los valores-p calculados para cada uno de estos tests sea válido? ¿Cuál es el valor de la devianza para el modelo ajustado?

Test de razón de Verosimilitud:
Las hipótesis testeadas en el test de razón de verosimilitud son las siguientes:

H_0 : Modelo restringido (nulo) $\theta = \theta_0$ H_1 : Modelo no restringido (con AGE) $\theta = \theta_1$

El estadístico asociado es:

$$D = -2log\left(\frac{\text{Log-verosimilitud Modelo restringido}}{\text{Log-verosimilitud Modelo no restringido}}\right)$$

Este test asume bajo la hipótesis nula una distribución asintótica del estadístico asociado D.

```
> modnul<-glm(barSTA ~ 1, family = binomial, data = datosUCI)
> library(mdscore)
> lr.test(modnul,model)

$LR
[1] 7.854589

$pvalue
[1] 0.005069187
```

```
attr(,"class")
[1] "lrt.test"

>
```

Utilizando una significancia del $\alpha = 5\%$, se concluye que el modelo nulo versus el modelo con la variable AGE poseen difernecias significativas respecto a la log-verosimilitud. Se rechaza que la inclusión de AGE al modelo nulo no incremente la log-verosimilitud de manera significativa.

Test de Wald:
Las hipótesis testeadas en el test de Wald son:

$$H_0 : \beta_{AGE} = 0 \quad H_1 : \beta_{AGE} \neq 0$$

El estadístico asociado a este test es:

$$\frac{\hat{\beta}^2_{AGE}}{var(\hat{\beta}_{AGE})}$$

Se asume que la diferencia $\hat{\beta} - \beta_0$ (en este caso β_0 es cero) distribuye normal y al dividirla por la desviación estándar de $\hat{\beta}$ se asumiría una distribución chi-cuadrado.

```
> wald.test(model, term=2)

$W
[1] 6.797355

$pvalue
[1] 0.009129303

attr(,"class")
[1] "wald.test"

>
```

Utilizando una significancia del $\alpha = 5\%$, se concluye que el parámetro β_{AGE} no podría asumirse nulo, en particular, la variable AGE sí tendría un efecto significativo en el modelo, pues el parámetro asociado a AGE sería no nulo bajo el test de Wald.

Test de Score:

Las hipótesis testeadas en el test de Wald son:

$$H_0 : R(\theta) = r \quad H_1 : R(\theta) \neq r$$

En este caso, considere R matriz de 2x2 con un 1 en (2,2) y r el vector nulo, con el fin de testear el coeficiente de AGE.

El estadístico asociado a este test es:

$$(R\hat{\theta}_n - r)^T [R(\hat{V}_n/n)R^T]^{-1}(R\hat{\theta}_n - r)$$

En el cual, dado que se asume que el vector $\hat{\theta}$ distribuye asintóticamente normal, el estadístico tendría distribución chi-cuadrado.

```
> 1-pchisq(mdscore(modnul, X=model.matrix(model)[,2])$Sr,
+          mdscore(modnul, X=model.matrix(model)[,2])$df)

[1] 0.007376774
```

Bajo el test de Score se rechaza la hipótesis de que el coeficiente β asociado a AGE sea nulo al 5%.

h) Usando los resultados de 1(e) calcule un intervalo de 95% de confianza para la pendiente y el intercepto. Interprete el intervalo de confianza para la pendiente.

```
> confint.default(model)

                 2.5 %      97.5 %
(Intercept)  1.69421908  4.42280739
AGE         -0.04824799 -0.00683723

> #Obtiene intervalo de confianza bajo normalidad asintótica
```

El intervalo de confianza a un 95% para la pendiente, es decir, β_1 (coeficiente asociado a AGE) es de (-0.048,-0.006), en el cual sólo se toman valores negativos, por lo que, podría aseverarse que efectivamente AGE se relacionaría negativamente con el log-odds de la sobrevivencia en la UCI (podría verse como un factor de riesgo el incremento de la edad de admisión). Si la edad de admisión aumenta en una unidad, se esperaría una disminución de entre -0.048 y -0.006 del log-odds o logaritmo de las chances de sobrevivencia.

i) Obtenga la matriz de varianzas-covarianzas estimada para el modelo ajustado en 1(e). Calcule el logito y la probabilidad estimada para una persona de 60 años de edad. Obtenga un intervalo de confianza de 95% para el logito y la probabilidad estimada. Interprete la probabilidad estimada y su intervalo de confianza.

La matriz de varianzas covarianzas estimada es la siguiente:

```
> vcov(model)

             (Intercept)           AGE
(Intercept)  0.484529087 -0.0071029945
AGE         -0.007102994  0.0001116015
```

El logito de sobrevivencia estimado para una persona de 60 años es:

$$log\left(\frac{P(STA=0|Edad=60)}{1-P(STA=0|Edad=60)}\right) = 3.058 - 0.027 \cdot 60 = 1.406$$

La probabilidad de sobrevivencia estimada para una persona de 60 años es:

$$\hat{\pi}_{60 \text{ años}} = \frac{exp(3.058 - 0.027 \cdot 60)}{1 + exp(3.058 - 0.027 \cdot 60)} = 0.803$$

Es decir, la probabilidad de sobrevivencia estimada para una persona de 60 años en la UCI es de 0.8 (en base a los datos y al modelo planteado).

Para obtener un intervalo de confianza del logito es posible utilizar:

```
> predicted = predict(model, data.frame(AGE=60), type='link', se.fit=TRUE)
> predicted$fit+c(-1,1)*predicted$se.fit

[1] 1.221741 1.590172
```

Y podemos obtener un intervalo de confianza usando propiedad de invarianza como sigue:

```
> exp(predicted$fit+
+     c(-1,1)*predicted$se.fit)/(1+exp(predicted$fit+c(-1,1)*predicted$se.fit))

[1] 0.7723698 0.8306403
```

Finalmente, podemos concluir que un intervalo de confianza para la probabilidad de sobrevivencia de un paciente de 60 años es de (0.772, 0.83) y un intervalo de confianza para el logaritmo de las chances de sobrevivencia para un paciente de 60 años es de (1.221, 1.59) note que en el intervalo toma valores mayores a 1, lo que indica que las chances de sobrevivir serían mayores que de no sobrevivir, al menos en el 95% de los casos.

j) Use el paquete de regresión logística para obtener el logito y su error estándar estimado para cada individuo en el estudio.

Para obtener el logito y desviación estándar de cada observación, se replica lo realizado anteriormente pero ahora a todas las observaciones:

```
> predictedall = predict(model, data.frame(AGE=AGE), type='link', se.fit=TRUE)
> head(predictedall$fit)     #Logito

        1         2         3         4         5         6
2.3148628 1.4334993 0.9377323 1.5712123 0.6623062 1.1580732

> head(predictedall$se.fit)  #Errores estándar de cada estimación

        1         2         3         4         5         6
0.4269952 0.1867103 0.2288089 0.2069681 0.3054835 0.1888142
```

2. Considere un modelo de regresión logística múltiple de STA sobre AGE, CAN, CPR, INF y RACE.

a) Escriba la ecuación para el modelo de regresión logística de STA sobre AGE, CAN, CPR, INF y RACE. Escriba la ecuación para la transformación logito de este modelo de regresión logística. ¿Cuántos parámetros contiene el modelo?

Notar que CAN, CPR, INF son variables de tipo factor de dos niveles y RACE de tres niveles. Luego, utilizando estas variables, en base al modelo planteado se tiene que $log\left(\frac{P(STA=0|AGE,CAN,CPR,INF,RACE)}{1-P(STA=0|AGE,CAN,CPR,INF,RACE)}\right)$ es igual a:

$$X\beta = \beta_0+\beta_1 AGE+\beta_2 \mathbb{1}_{CAN=Si}+\beta_3 \mathbb{1}_{CPR=Si}+\beta_4 \mathbb{1}_{INF=Si}+\beta_5 \mathbb{1}_{RAC=Negro}+\beta_6 \mathbb{1}_{RAC=Otra}$$

Y la probabilidad se puede expresar como:

$$\pi_i = \frac{exp(X\beta_i)}{1+exp(X\beta_i)}$$

El modelo contiene 7 parámetros:

- El intercepto
- 1 coeficiente asociado a CAN con celda de referencia CAN=No
- 1 coeficiente asociado a CPR con celda de referencia CPR=No
- 1 coeficiente asociado a INF con celda de referencia INF=No
- 2 coeficientes asociados a RACE con celda de referencia RACE=Blanco

b) Usando un paquete de regresión logística obtenga las estimaciones de máxima verosimilitud para los parámetros del modelo de regresión logística de STA sobre AGE, CAN, CPR, INF y RACE. Escriba la ecuación para la transformación logito de este modelo de regresión logística y la ecuación para los valores ajustados, esto es, las probabilidades logísticas estimadas.

Para plantear el modelo es necesario definir las variables de tipo factor:

```
> datosUCI$CANf<-factor(ifelse(CAN=="2", "Si", "No"))
> datosUCI$CPRf<-factor(ifelse(CPR=="2", "Si", "No"))
> datosUCI$INFf<-factor(ifelse(INF=="2", "Si", "No"))
> datosUCI$RACf<-factor(ifelse(RAC=="1", "Blanco",
+                              ifelse(RAC=="2", "Negro", "Otra")))
> model2<-glm(barSTA ~ AGE+CANf+CPRf+INFf+RACf,
+             family = binomial, data = datosUCI)
> coef(model2)

(Intercept)         AGE      CANfSi      CPRfSi      INFfSi   RACfNegro
 3.51151970 -0.02712074 -0.24451057 -1.64649699 -0.68066763  0.95707769
   RACfOtra
-0.25974927
```

Los parámetros se obtuvieron maximizando la log-verosimilitud utilizando el método de Fisher Scoring, el cual convergió en 5 iteraciones. Luego, se puede establecer que en el modelo el logito estimado es igual a:

$$X\hat{\beta} =$$

$$3.51 - 0.02 \cdot AGE - 0.24 \cdot 1\!\!1_{CAN=Si} - 1.64 \cdot 1\!\!1_{CPR=Si} - 0.68 \cdot 1\!\!1_{INF=Si} + 0.95 \cdot 1\!\!1_{RAC=Negro} - 0.25 \cdot 1\!\!1_{RAC=Otra}$$

Y la probabilidad estimada:

$$\hat{\pi}_i = \frac{exp(X\hat{\beta})_i}{1 + exp(X\hat{\beta})_i}$$

c) Evalúe la significancia de los coeficientes del modelo usando el test de razón de verosimilitud. ¿Qué supuestos son necesarios para que los valores-p calculados para estos test sean válidos? ¿cuál es el valor de la devianza para el modelo ajustado?

```
> lr.test(modnul,model2)

$LR
[1] 20.86024

$pvalue
[1] 0.001943745

attr(,"class")
[1] "lrt.test"

>
```

Utilizando una significancia del $\alpha = 5\%$, se concluye que el modelo nulo versus el modelo con las variables AGE, CAN, CPR, INF y RACE, poseen diferencias significativas respecto a la bondad de ajuste bajo el test de verosimilitud. Se rechaza que la inclusión de las variables antes mencionadas al modelo nulo no mejore significativamente la bondad de ajuste del modelo nulo.

Este test asume bajo la hipótesis nula una distribución asintótica del estadístico asociado $D = 2log\lambda$ donde λ es el cuociente de la verosimilitud del modelo nulo entre la verosimilitud del modelo de interés.

La devianza del modelo ajustado es:

```
> deviance(model2)

[1] 179.3007
```

También es posible estudiar la inclusión de cada variable (AGE ya se estudió en el ítem anterior así que no se volverá a analizar):

```
> modCANf<-glm(barSTA ~ CANf, family = binomial, data = datosUCI)
> modcPRf<-glm(barSTA ~ CPRf, family = binomial, data = datosUCI)
> modINff<-glm(barSTA ~ INFf, family = binomial, data = datosUCI)
> modRACf<-glm(barSTA ~ RACf, family = binomial, data = datosUCI)
> lr.test(modnul,modCANf)$pvalue

[1] 0.9999999

> lr.test(modnul,modcPRf)$pvalue

[1] 0.004855682

> lr.test(modnul,modINff)$pvalue

[1] 0.01033844

> lr.test(modnul,modRACf)$pvalue

[1] 0.3230539

>
```

Podemos observar que las variables que en el test de razón de verosimilitud no resultas aportar bondad de ajuste respecto al modelo nulo es la variable CAN (valor-p aproximadamente 1), la que indicaba si el paciente presentaba problemas o no de cáncer, lo anterior es un hallazgo bastante interesante, pues estamos estudiando sobrevivencia y la variable RACE la que indicaba la raza del paciente.

Estos hallazgos también se complementan con los tests de significancia usuales:

```
> summary(model2)

Call:
glm(formula = barSTA ~ AGE + CANf + CPRf + INFf + RACf, family = binomial,
    data = datosUCI)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5124   0.3082   0.5421   0.6823   1.3703

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.51152    0.81443   4.312 1.62e-05 ***
AGE         -0.02712    0.01159  -2.340  0.01926 *
CANfSi      -0.24451    0.61681  -0.396  0.69180
CPRfSi      -1.64650    0.62341  -2.641  0.00826 **
INFfSi      -0.68067    0.38042  -1.789  0.07357 .
RACfNegro    0.95708    1.08445   0.883  0.37748
RACfOtra    -0.25975    0.87127  -0.298  0.76561
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 200.16  on 199  degrees of freedom
Residual deviance: 179.30  on 193  degrees of freedom
AIC: 193.3

Number of Fisher Scoring iterations: 5
```

d) Use el estadístico de Wald para obtener una aproximación a la significancia individual de los coeficientes para las variables en el modelo. Ajuste un modelo reducido que elimine aquellas variables no significativas. Chequee la significancia conjunta (condicional) de las variables excluídas del modelo. Presente los resultados del ajuste para el modelo reducido en una tabla.

```
> wald.test(model2, term=3)$pvalue  #Variable CAN

[1] 0.6918037

> wald.test(model2, term=4)$pvalue  #Variable CPR

[1] 0.008263709

> wald.test(model2, term=5)$pvalue  #Variable INF
```

```
[1] 0.07357212

> wald.test(model2, term=6)$pvalue  #Variable RAC Negro

[1] 0.3774835

> wald.test(model2, term=7)$pvalue  #Variable RAC Otra

[1] 0.7656056

>
>
```

En base al test de Wald, es posible observar que las variables que no resultan ser significativas son CAN y RAC, además la variable INF posee un valor-p de 0.07, si se utiliza un 5% de significancia también quedaría fuera aunque sería disctuble. Se ajusta el modelo sin estas tres variables:

```
> modsign<-glm(barSTA ~ AGE+CPRf, family = binomial, data = datosUCI)
> #Testea la diferencia de la devianza del modelo completo y reducido:
>
> anova(modsign, model2)

Analysis of Deviance Table

Model 1: barSTA ~ AGE + CPRf
Model 2: barSTA ~ AGE + CANf + CPRf + INFf + RACf
  Resid. Df Resid. Dev Df Deviance
1       197     183.95
2       193     179.30  4    4.652

> AIC(modsign,model2)

        df      AIC
modsign  3 189.9527
model2   7 193.3007

>
> #Quitando las 3 variables no significativas, no se observa una pérdida
> # muy grande en términos de bondad de ajuste

% latex table generated in R 4.0.5 by xtable 1.8-4 package
% Thu Jun 10 14:41:36 2021
\begin{table}[ht]
\centering
```

```
\begin{tabular}{rrrrr}
  \hline
 & Estimate & Std. Error & z value & Pr($>$$|$z$|$) \\
  \hline
(Intercept) & 3.3520 & 0.7455 & 4.50 & 0.0000 \\
  AGE & -0.0296 & 0.0111 & -2.66 & 0.0079 \\
  CPRfSi & -1.7841 & 0.6073 & -2.94 & 0.0033 \\
   \hline
\end{tabular}
\end{table}
```

Los resultados del modelo reducido se presentan a continuación:

	Estimate	Std. Error	z value	Pr(>|z|)
Intercepto	3.3520	0.7455	4.50	0.0000
Edad	-0.0296	0.0111	-2.66	0.0079
Resucitación pulmonar Sí	-1.7841	0.6073	-2.94	0.0033

Y el análisis de devianza para las componetnes del modelo:

```
% latex table generated in R 4.0.5 by xtable 1.8-4 package
% Thu Jun 10 14:41:36 2021
\begin{table}[ht]
\centering
\begin{tabular}{lrrrr}
  \hline
 & Df & Deviance & Resid. Df & Resid. Dev \\
  \hline
NULL &  &  & 199 & 200.16 \\
  AGE & 1 & 7.85 & 198 & 192.31 \\
  CPRf & 1 & 8.35 & 197 & 183.95 \\
   \hline
\end{tabular}
\end{table}
```

	Df	Deviance	Resid. Df	Resid. Dev
Modelo nulo			199	200.16
Modelo con Edad	1	7.85	198	192.31
Modelo con resucitación pulmonar	1	8.35	197	183.95

3. Considere ahora como variable respuesta a STA y a CPR como covariable.

a) Demuestre que el valor del log-odds ratio obtenido de la clasificación cruzada de STA y CPR es idéntico a la pendiente estimada para la regresión logística de STA

sobre CPR. Verifique que los errores estándar estimados para el coeficiente de CPR es idéntico a la raíz cuadrada de la suma de los inversos de las frecuencias en las celdas de la tabla de clasificación cruzada de STA por CPR. Obtenga un intervalo de 95% de confianza para el odds ratio. ¿Qué característica de la variable CPR hace equivalentes los cálculos para los dos métodos?

```
> modeloCPR<-glm(barSTA ~ CPRf, family = binomial,
+                data = datosUCI)
> coef(modeloCPR)

(Intercept)      CPRfSi
   1.540445   -1.694596
```

log-odds ratio obtenido de la clasificación cruzada de STA y CPR es idéntico a la pendiente estimada para la regresión logística de STA sobre CPR

El logito en este modelo es:

$$log\left(\frac{P(STA=0|CPR)}{1-P(STA=0|CPR)}\right) = 1.540445 - 1.6945961\mathbb{I}_{CPR=Si}$$

De donde:

$$OR(CPR = Si \quad \text{respecto} \quad CPR = No) = e^{-1.694596} = 0.1836734$$

Las chances de sobrevivir al haber tenido resucitación pulmonar son aproximadamente 5 veces menores respecto a una persona que no necesitó resucitación pulmonar. Podríamos decir, que haber necesitado resucitación pulmonar es un factor de riesgo para la sobrevivencia.

La clasificación cruzada es:

```
> table(datosUCI$barSTA, datosUCI$CPRf)

    No  Si
  0  33   7
  1 154   6
```

Notar que, efectivamente, el odds Ratio calculado coincide con la pendiente del modelo:

$$OR(CPR = Si \quad \text{respecto} \quad CPR = No) = \frac{33 \cdot 6}{154 \cdot 7} = 0.1836735$$

Errores estándar estimados para el coeficiente de CPR es idéntico a la raíz cuadrada de la suma de los inversos de las frecuencias en las celdas de la tabla de clasificación

cruzada de STA por CPR

```
> sqrt(vcov(modeloCPR)[2,2])  #Error estandar pendiente asociada a CPR

[1] 0.5884899
```

Y notar que:

$$\sqrt{\frac{1}{33}+\frac{1}{7}+\frac{1}{154}+\frac{1}{6}} = 0.5884898$$

Equivalente al error estándar de la pendiente asociada al modelo.

Obtenga un intervalo de 95% de confianza para el odds ratio

Utilizando el error estándar y la estimación de la pendiente de CPR obtenemos un intervalo de confianza del coeficiente asociado a CPR:

```
> coef(modeloCPR)[2] +c(-1,1)*sqrt(vcov(modeloCPR)[2,2])

[1] -2.283086 -1.106106
```

Utilizando invarianza, podemos obtener un intervalo para el odds ratio:

```
> exp(coef(modeloCPR)[2] +
+       c(-1,1)*sqrt(vcov(modeloCPR)[2,2]))

[1] 0.1019691 0.3308448
```

¿Qué característica de la variable CPR hace equivalentes los cálculos para los dos métodos?

La variable CPR es una variable de tipo dicotómica, es por esto que hacer los cálculos con la tabla cruzada se pueden equiparar a los coeficientes obtenidos en el modelo de regresión logística.