PLANTEO DEL PROBLEMA

Suponga que se desea estudiar una enfermedad en relación a tres variables que se presume tienen que ver con la ocurrencia de aquella (el sexo, ser o no fumador, y la edad).
Para ello se interroga a 7 individuos sanos (libres de la enfermedad) al comienzo del estudio, recabando información sobre las tres variables mencionadas. Luego de transcurridos 10 años se los contacta nuevamente, evaluando quienes han desarrollados la enfermedad.
Sin duda se trata de un estudio de seguimiento o follow-up

Los resultados fueron:

AL INICIO			10 AÑOS DESPUES
SEXO	**FUMA**	**EDAD**	**ESTADO**
M	NO	40	E
F	SI	70	S
M	NO	55	S
F	SI	60	E
M	SI	30	E
F	NO	35	S
M	SI	45	S

La intención es buscar una fórmula (ecuación de regresión) que permita, en base a la variables del individuo al principio del estudio, predecir su estado 10 años después

PREPARACIÓN DE LOS DATOS

Pero antes necesitamos que las variables del individuo al principio del estudio, o variables predictoras, sean numéricas.
Entonces las recodificamos como sigue, y para emplear una notación más concisa les cambiaremos el nombre:

SEXO, será X, con valores 0 (F) 1 (M)
FUMA, será Y, con valores 0 (NO) 1 (SI)
EDAD, será Z, respetando sus valores

Lugo el archivo de datos será:

X	**Y**	**Z**	**ESTADO**
1	0	40	E
0	1	70	S
1	0	55	S
0	1	60	E
1	1	30	E
0	0	35	S
1	1	45	S

PORQUE REGRESIÓN LOGÍSTICA

Las variables predictoras son X, Y, Z y la variable de salida es ESTADO.
Notar que la variable de salida, ESTADO, es dicotómica, admite solo dos valores ENFERMO o SANO y es justamente esta característica de la variable de salida, el ser dicotómica, lo que hace necesario emplear Regresión Logística.
Por ejemplo, si en este estudio la variable de salida de interés hubiese sido Nivel de Colesterol, que es continua, se podría haber empleado regresión lineal múltiple para predecir el Nivel de Colesterol en base a estas mismas variables X, Y, Z. En este caso lo que se predice es claro, es el Nivel de Colesterol del individuo 10 años después (que puede diferir mucho o poco del Nivel real medido).
Pero en nuestro caso de RL, que es lo que se quiere predecir?.
El estado, pero a través de una probabilidad.

FORMULA DE PREDICCION

Pero que fórmula usaremos para predecir el estado de un individuo (E o S) en base a X, Y y Z?.

En regresión Logística se usa

$$P(E) = \frac{1^{-H}}{1+e^{-H}}$$

en donde H es una función con coeficientes constantes de las variables X, Y y Z.
Por ejemplo podríamos tomar:

$H = \alpha + \beta X$
o $H = \alpha + \beta_1 X + \beta_2 Y + \beta_3 Z$
o $H = \alpha + \beta_1 X + \beta_2 Y + \gamma X Y$
o $H = \alpha + \beta_1 X + \beta_2 X^2 + \beta_3 Y$
etc.

En regresión logística, la función H empleada, define el "modelo" a ajustar.
En el resto del libro se indicará que modelo conviene utilizar en cada caso.
En lo que sigue, supongamos que el "modelo elegido" es:

$$H = \alpha + \beta_1 X + \beta_2 Y + \beta_3 Z$$

Luego la fórmula de predicción que proponemos aquí es:

$$P(E) = \frac{1}{1 + e^{-(\alpha + \beta_1 X + \beta_2 Y + \beta_3 Z)}}$$

De acuerdo, esta es una "fórmula tipo" pero para poder emplearla para predecir el estado de

enfermedad futuro del individuo, necesitamos conocer α, β_1, β_2, y β_3.

METODO DE ESTIMACIÓN DE MÁXIMA VEROSIMILITUD

Estos valores les calcula el soft de RL, utilizando un método para estimarlos en base a los datos, llamado método de máxima verosimilitud.
Veamos en que consiste:

1º) El soft elige unos valores iniciales para α, β_1, β_2 y β_3 por ejemplo

$\hat{\alpha} = -1{,}5$ $\quad \hat{\beta}_1 = 1{,}6$ $\quad \hat{\beta}_2 = -2{,}8$ $\quad \hat{\beta}_3 = 0{,}04$

Luego la fórmula queda

$$P(E) = \frac{1}{1 + e^{-(-1{,}5 + 1{,}6\,X - 2{,}8\,Y + 0{,}04\,Z)}}$$

2º) seguidamente, para evaluar que tan bien predice esta fórmula, el soft la aplica a cada uno de los 7 individuos del estudio.
Para el individuo 1: (X = 1 Y = 0 Z = 40) que enfermó

$$P(E_1) = \frac{1}{1 + e^{-(-1{,}5 + 1{,}6 * 1 - 2{,}8 * 0 + 0{,}04 * 40)}} = 0{,}85$$

Para el individuo 2: (X = 0 Y = 1 Z = 70) que se mantiene sano

$$P(S_2) = 1 - P(E_2) = 1 - \frac{1}{1 + e^{-(-1{,}5 + 1{,}6 * 0 - 2{,}8 * 1 + 0{,}04 * 70)}} = 0{,}82$$

Para el individuo 3: (X = 1 Y = 0 Z = 55) se mantiene sano

$$P(S_3) = 1 - P(E_3) = 1 - \frac{1}{1 + e^{-(-1{,}5 + 1{,}6 * 1 - 2{,}8 * 0 + 0{,}04 * 55)}} = 0{,}09$$

Siguiendo con los cálculos resulta al final
$P(E_1) = 0{,}85$
$P(S_2) = 0{,}82$
$P(S_3) = 0{,}09$
$P(E_4) = 0{,}13$
$P(E_5) = 0{,}18$

$P(S_6) = 0{,}52$
$P(S_7) = 0{,}71$

Con el individuo 1 la fórmula funciona muy bien, ya que predice en base a X, Y y Z una probabilidad de enfermar de 0,85 y el individuo realmente enfermó.
Con el individuo 2 también, ya que predice una probabilidad de no enfermar de 0,82 y el individuo realmente no enfermó.
Sin embargo con el individuo 3 la predicción fue muy mala ya que predice una probabilidad de no enfermar o de estar sano muy baja, de 0,09 cuando en realidad el individuo está sano.
También es mala la predicción en los individuo 4, 5 y 6 siendo buena en el 7^{mo}.

3º) Como se vio, está fórmula en algunos individuos predice bien su estado de enfermedad, y en otros no. Lo ideal sería que las probabilidades predichas en [2] sean todos números cercanos a 1.

Para evaluar globalmente la bondad de predicción de la fórmula, necesitamos un indicador.
El método de máxima verosimilitud que estamos describiendo, utiliza como indicador de bondad de predicción, el producto de las probabilidades de predicción

$L = P(E_1) * P(S_2) * P(S_3) * P(E_4) * P(E_5) * P(S_6) * P(S_6) = 0{,}000556$

A esta cantidad se le llama verosimilitud (del ingles likelihood)

Mide la calidad con que la fórmula predice el estado de salud final de los 7 individuos en base a sus variables X, Y y Z.

Notar que si la predicción es perfecta
$P(E_1) = 1$, $P(S_2) = 1$,… $P(E_7) = 1$
Y entonces, $L = 1 * 1 * 1 * ... 1* = 1$

En cambio, en el peor caso de predicción
$P(E_1) = 0$, $P(S_2) = 0$… $P(E_7) = 0$
Y entonces, $L = 0 * 0 * 0 * ... * 0 = 0$

O sea la verosimilitud L, siempre está entre 0 y 1
En este caso fue $L = 0{,}000556$

4º) Ahora el soft vuelve al paso 1º) pero eligiendo otros valores para α, β_1, β_2, y β_3.
Por ejemplo:

$\hat{\alpha} = -2$ $\hat{\beta}_1 = 1{,}8$ $\hat{\beta}_2 = -3$ $\hat{\beta}_3 = 0{,}05$

y repite los pasos 2º) y 3º).

Si ahora la verosimilitud L, aumentó, quiere decir que esta nueva fórmula es mejor, pues

predice mejor el estado de enfermedad de los 7 individuos de la muestra.
En el caso que disminuya, se descarta la fórmula, ya que con ella la predicción empeora.

El soft repite continuamente este proceso de prueba (pasos 1-2 y 3) hasta que finalmente obtiene unos coeficientes $\hat{\alpha}$, $\hat{\beta}_1$, $\hat{\beta}_2$ y $\hat{\beta}_3$ que dan la más alta verosimilitud L. L = 0,01496

$\hat{\alpha}$ = 1,78 $\hat{\beta}_1$= + 0,168 $\hat{\beta}_2$ = + 1,252 $\hat{\beta}_3$ = - 0,061

Con estos coeficientes, la fórmula predice de la mejor manera (con máxima verosimilitud L) el estado de enfermedad de los individuos de la muestra [3]

$$P(ENF) = \frac{1}{1 + e^{-(1,78 + 0,168 * X + 1,252 * Y - 0,061 * Z)}}$$

Resumiendo, cuando se termina el proceso de iteraciones del soft, la salida que proporciona son:

a) Las estimaciones $\hat{\alpha}$, $\hat{\beta}_1$, $\hat{\beta}_2$ y $\hat{\beta}_3$
b) La verosimilitud máxima obtenida L

OBSERVACIONES RESPECTO DEL SOFT

Si el soft repitiese las pasos mencionados eligiendo arbitrariamente las estimaciones del paso 1º, tardaría muchísismo tiempo en lograr las estimaciones que den la verosimilitud máxima.
Por ello, en cada paso 1, el soft tiene en cuenta las estimaciones anteriores, y las verosimilitudes obtenidas, proponiendo nuevas estimaciones, que presumiblemente incrementen la verosimilitud.
De esta manera, en menor número de iteraciones se logran descubrir las estimaciones $\hat{\alpha}$, $\hat{\beta}_1$, $\hat{\beta}_2$ y $\hat{\beta}_3$ que producen la verosimilitud máxima.
Dicho en lenguaje matemático, se logra la convergencia del proceso iterativo.

UTILIDAD DEL MODELO: PREDICCION

Hemos ajustado un modelo logrando una fórmula de predicción, que predice de la mejor manera, el estado de enfermedad de los 7 individuos del estudio.
Pero ahora queremos utilizar dicha fórmula para predecir el estado de enfermedad futura, de otro individuo, que no pertenece a esta muestra, por ejemplo una mujer, fumadora, que hoy tiene 31 años, o sea con variables predoctoras

X = 0 Y = 1 Z = 31

Si aplicamos la fórmula [3]:

$$P(ENF) = \frac{1}{1 + e^{-(1{,}78 + 0{,}168 * 0 + 1{,}252 * 1 - 0{,}061 * 31)}} = 0{,}76$$

Por supuesto ahora no tenemos forma de evaluar que tan buena es esta predicción, ya que no sabemos como evolucionará está mujer en las próximos 10 años.
Se trata realmente de una predicción de su estado de enfermedad futuro.
Lo que si sabemos es que esta formula de predicción funcionó razonablemente bien con los 7 individuos de la muestra, en los que si conocíamos su estado de enfermedad a los 10 años.

Y esto fue así, porque expresamente buscamos con el método de máxima verosimilitud, obtener una fórmula que lo mejor posible, logre predecir el estado de los individuos de la muestra.
Por eso, cuando el objetivo principal al efectuar una RL es predicción en individuos nuevos, es deseable que el ajuste en la muestra sea muy bueno y esto equivale a lograr una verosimilitud máxima alta.
Por ejemplo en nuestro caso, al proponer el modelo de ajuste

$$H = \alpha + \beta_{1,} X + \beta_{2,} Y + \beta_3 Z$$

Logramos una verosimilitud máxima de $L = 0{,}01496$

Pero con este modelo no podemos superar $L = 0{,}01496$ ya que es la máxima, pero quizás, si hubiésemos propuesto otro modelo como:

$$H = \alpha + \beta_{1,} X + \beta_{2,} Y + \beta_3 Z + \beta_4 XY$$

La verosimilitud máxima pudo haber sido $L = 0{,}032$
Si este fuese el caso, con este modelo, la predicción de la PROB (ENFERMAR) de la mujer de 31 años hubiese sido mejor.

En definitiva se quiere resaltar que si el objetivo principal de una RL es predicción, debemos preocuparnos en lograr un modelo que ajuste con una verosimilitud máxima, lo mayor posible.
Posteriormente se comentarán objeciones a este proceder, basados en el principio de parsimonia.

OUTPUT DE UN SOFT DE REGRESIÓN LOGÍSTICA

A continuación se comentarán solo las salidas del soft vinculados con los temas tratados.
En nuestro caso, si el modelo propuesto para el ajuste es:

$$[4]\ H = \alpha + \beta_{1,} X + \beta_{2,} Y + \beta_3 Z$$

La salida incluirá la tabla

	B	**S.E.**	**Sig.**
X	0,168	1,849	0,928
Y	1,252	1,818	0,491
S	-0,061	0,077	0,433
Constante	1,780	4,046	0,660

-2 Log. Likelihood = 8,404

En la columna encabezada por B, están los coeficientes estimados del modelo propuesto, que llamamos $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ y $\hat{\alpha}$.
En la siguiente columna, encabezada por S.E. están los errores estándar de estos coeficientes. Y en la última figuran los NIVELES de significación p en las pruebas de hipótesis de cada coeficiente del modelo, respecto de cero.
En nuestro ejemplo no se puede rechazar la hipótesis nula H_0: $\beta_1 = 0$, ya que $p = 0{,}928 > 0{,}05$.
También pasa lo mismo con H_0: $\beta_2 = 0$, ya que $p = 0{,}491 > 0{,}05$ y similarmente para los restantes coeficientes.
O sea no hay evidencia muestral que contradiga la suposición de que los coeficientes del modelo [4] son nulos.
Pero esto quizás sea así debido al pequeño tamaño de muestra (n = 7) de este ejemplo.
Y otra consecuencia de este tamaño muestral, son los relativamente grandes errores estándar de la columna 2, respecto de los coeficientes estimados de la columna 1.
Otra forma de averiguar la significación de un coeficiente sería utilizando el estadístico normal z.

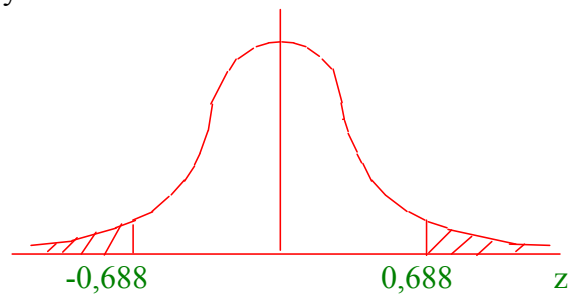
$$z = \frac{B}{S.E.} \rightarrow N(0;1)$$

Por ejemplo para testear

$$\begin{cases} H_0: & \beta_2 = 0 \\ H_0: & \beta_2 \neq 0 \end{cases}$$

Se calcula $z = \frac{B}{S.E.} = \frac{1{,}252}{1{,}818} = 0{,}688$

Y yendo a una tabla normal



Resulta del área sombreada, p = 0,491, No significativo

Test de WALD

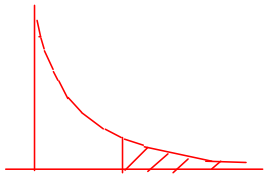
Teniendo en cuenta que el cuadrado de una variable normal estándar se distribuye según una χ^2_1 resulta

$$\left(\frac{B}{S.E.}\right)^2 \rightarrow \chi_1^2$$

El test anterior puede efectuarse calculando

$$\left(\frac{B}{S.E.}\right)^2 = \left(\frac{1,252}{1,818}\right)^2 = (0,688)^2 = 0,474$$

Y yendo a una tabla Chi-cuadrado con 1 grado de libertad



0,474 χ_1^2

Resultando del área sombreada el mismo p = 0,491

Queda un último comentario respecto de la verosimilitud máxima con que se logra el ajuste del modelo.
En nuestro ejemplo fue L = 0, 01496, sin embargo en general los softs no proporcionan esta verosimilitud.
Le calculan el logaritmo natural, multiplican el resultado por – 2, e informan esta cantidad

$$\ln L = \ln (0,01496) = - 4,2023$$

Luego

$$- 2 \ln L = (-2) * (- 4,2023) = - 8,404$$

Pero porque informan –2 en L en lugar de L?
El motivo es que – 2 ln L tiene propiedades estadísticas especiales, que permiten comparar entre si el ajuste de distintos modelos. Pero esto lo veremos más adelante.
Sin embargo L y – 2 ln L están relacionadas. Recordemos que si el ajuste es perfecto

$$L = 1$$

Entonces – 2 ln L = - 2 ln (1) = - 2 * 0 = 0
En cambio si el ajuste es muy malo L ≈ 0 resulta
– 2 ln L = - 2 ln (0) → ∞

O sea resumiendo:
Buen ajuste: L alto y - 2 ln L bajo
Mal Ajuste: L bajo y - 2 ln L alto

Esto significa que podemos interpretar a $-2 \ln L$ como una medida del error del ajuste; si el ajuste es muy bueno su error de ajuste, $-2 \ln L$ será cercano a cero; en cambio si el ajuste es malo, tendrá mucho error, y por tanto $-2 \ln L$ será alto.

USO DEL MODELO II: EVALUACIÓN DE VARIABLES

Anteriormente utilizamos el modelo de regresión con fines predictivos. Concretamente para predecir el estado, diez años después, de un individuo no perteneciente a la muestra y vimos que para lograr predicciones confiables en general es deseable un modelo que ajuste bien en la muestra, o sea con L alta, o lo que es equivalente con baja $-2 \ln L$.

Pero en lo que sigue utilizaremos el modelo de regresión con otro fin, concretamente para evaluar variables de riesgo.
Por ejemplo queremos conocer:

- Los que fuman, cuanto más riesgo de enfermar tienen, respecto de los que no fuman?
- Y los varones respecto de las mujeres?
- O en cuanto aumenta el riesgo de enfermar por cada año?

Para dar respuestas a estas preguntas debemos interpretar los coeficientes del modelo ajustado.
Y como dicha interpretación varía según el modelo, le dedicaremos varios capítulo a este tema.

Por último, en este enfoque, si bien un buen ajuste del modelo a los datos es importante, le prestaremos mucha atención a aspecto como la validez del modelo, y a los errores estándar de las estimaciones de los coeficientes.

MEDIDAS DE RIESGO EN REGRESIÓN LOGÍSTICA: ODDS y LOGIT

Anteriormente hemos comentado que para medir el "riesgo de enfermar" de un individuo lo más natural es utilizar la "probabilidad de enfermar" o sea P(E).
Sin embargo, otra forma válida de medir el "riesgo de enfermar", aunque no tan difundido es a través del "ODDS de enfermar"

$$ODDS_E = \frac{P(E)}{1-P(E)}$$

Se trata de dos medidas diferentes del mismo concepto: "riesgo de enfermar".

Por ejemplo, si respecto de cierta enfermedad, la "probabilidad de enfermar" de un individuo es $P(E) = 0,6$.

Entonces su "ODDS de enfermar" es $ODDS_E = \frac{P(E)}{1-P(E)} = \frac{0,6}{1-0,6} = 1,5$

Dicho de otra forma, en cuanto a riesgo a "riesgo de enfermar" es equivalente tener
$P(E) = 0,6$ o $ODDS_E = 1,5$

Será sin embargo el problema que estemos considerando, el que aconseje utilizar P(E) o $ODDS_E$ para medir el "riesgo de enfermar".

Un ejemplo de la conveniencia de pensar con ODDS, es el de la Regresión Logística:
Pero veamos porqué y en donde.

Cuando se emplea RL para evaluar la asociación entre algunas variables X_1 X_k, y una enfermedad E, lo primero que hay que proponer es el modelo a ajustar.
Por ejemplo

$$H = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_k X_k$$

Y el soft, utilizando la fórmula

$$P(E) = \frac{1}{1+e^{-H}}$$

O sea $$P(E) = \frac{1}{1 + e^{-(\alpha + \beta_1 X_1 + \beta_k X_k)}}$$

Calcula la verosimilitud de la muestra

$$L = P(E_1) * P(S_2) * P(E_7) ***** P(E_n)$$

Y estima los coeficientes $\hat{\alpha}$, $\hat{\beta}_1$, $\hat{\beta}_k$, que hacen máxima esta verosimilitud.
En esta etapa el soft evalua el "riesgo de enfermar" con probabilidades $P(E_i)$, ya que son probabilidades, las que aparecen de la verosimilitud L.

Pero veamos quien es H. Despegando de [2] resulta:

$$1+e^{-H} = \frac{1}{P(E)}$$

$$e^{-H} = \frac{1}{P(E)} - 1 = \frac{1 - P(E)}{P(E)}$$

Como $e^{-H} = \frac{1}{e^H}$ resulta

$$\frac{1}{e^H} = \frac{1 - P(E)}{P(E)}$$

$$e^H = \frac{P(E)}{}$$

1-P(E)

Y calculando en ambos miembros de esta expresión

$$\ln (e^{H}) = \ln \left(\frac{P(E)}{1-P(E)} \right)$$

Pero: $\ln (e^{H}) = H$

$$Ln \left(\frac{P(E)}{1-P(E)} \right) = \ln ODDS(E)$$

quedan $\boxed{H = \ln ODDS(E)}$

Al depender solo de ODDS(E), H es otra medida del "riesgo de enfermar".
Se llama a esta medida del riesgo, el "LOGIT de ENFERMAR", Logit(E).

Siguiendo con el ejemplo del principio de esta sección, en cuanto a "riesgo de enfermar" es equivalente decir que un individuo tiene:

$P(E) = 0,6$
o $ODDS(E) = 1,5$
o $LOGIT(E) = \ln ODDS(E) = \ln (1,5) = 0,405$

Se trata de tres medidas diferentes en diferente escalas, de el mismo "riesgo de enfermar"
Volviendo a nuestro modelo de RL, otra forma de expresarlo es (ver [1])

$$\boxed{\ln ODDS (E) = \alpha + \beta_1 X_1 + \beta_k X_k}$$

En donde se pone de manifiesto, que lo que evalúa la RL con la expresión
$\alpha + \beta_1 X_1 + \beta_k X_k$ es el "riesgo de enfermar", pero medido en logits.

En nuestro ejemplo de este capítulo, el modelo ajustado fue:

$$\ln ODDS (E) = 1,780 + 0,168\, X + 1,252\, Y - 0,061\, Z$$

Entonces para un individuo masculino (X = 1), que no fuma (Y = 0) y tiene (Z = 42) 42 años, su "riesgo de enfermar", medido en logits es:

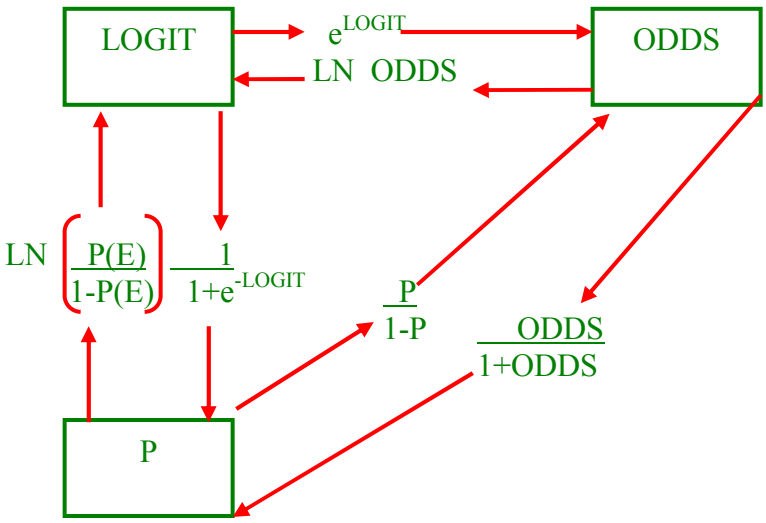
$$\ln ODDS (E) = 1,780 + 0,168\, 1 + 1,252\, 0 - 0,061\, 42 = -0,614$$

Y en medido en ODDS es: $ODDS (E) = e^{-0,614} = 0,541$

Y medio en probabilidad es:

$$P(E) = \frac{1}{1 + e^{-(-0,614)}} = 0,351$$

EQUIVALENCIA ENTRE MEDIDAS DE RIESGO



COMPARACIÓN DE RIESGOS EN REGRESIÓN LOGÍSTICA: OR y RR

En el apartado anterior analizamos como medir riesgos en Regresión Logística, concluyendo que es equivalente hacerlo con una probabilidad de enfermar, PROB (E), un odds ODDS (E) o sin logit, LOGIT (E).
Ahora estudiaremos como comparar riesgos:
Se define PERFIL DE RIESGO de un individuo, a el valor que toman sus variables predictoras.
Por ejemplo, en el caso que estamos analizando una mujer, fumadora de 37 años tiene un perfil de riesgo caracterizado por:

$$X = 0 \qquad Y = 1 \qquad Z = 37$$

En lo que sigue la intención es comparar riesgos entre individuos que tienen diferentes perfiles de riesgo, o sea que difieren en sus variables predictoras.

Cuando un perfil de riesgo especifica completamente el valor de TODAS las variables predictoras, el modelo asigna un único riesgo para todos los individuos con este perfil.
Por ejemplo para el perfil:

$$X = 0 \qquad Y = 1 \qquad Z = 37$$

$$Logit\ (E) = 1,78 + 0,168 * 0 + 1,252 * 1 - 0,061 * 37 = 0,775$$

$$ODDS\ (E) = e^{0,775} = 2,17$$

$$PROB\ (E) = \frac{ODDS\ (E)}{1+ODDS\ (E)} = \frac{2,17}{1+2,17} = 0,68$$

Entonces cualquier mujer, fumadora de 37 años tiene luego de 10 años, una probabilidad de estar enferma de 0,68, o lo que es lo mismo un ODDS de estar enferma de 2,17, o un Logit de 0,775.
Por supuesto estas son estimaciones de riesgo, proporcionadas por el modelo.
Pero a veces, por motivos que estudiaremos en este apartado, suele ser de interés analizar perfiles de riesgo que dejan sin especificar algunas variables predictoras.
Por ejemplo el perfil

$X = 0 \quad Y = 1 \quad Z = ?$

O también escrito así

$X = 0 \quad Y = 1 \quad Z = z$

Este perfil caracteriza a todas las mujeres, fumadoras, independientemente de su edad, o sea que pertenece a este perfil tanto una mujer fumadora de 37 años, como la que tiene 14 o 82, o cualquier otra edad.
Pero en estos casos de perfiles incompletas, el riesgo que asigna el modelo, no es el mismo para todos los individuos del perfil, concretamente en este caso:

$$Logit\ (E) = 1,78 + 0,168 * 0 + 1,252 * 1 - 0,061 * z$$

Resultando para $z = 37$ años un $Logit = 0,775$, para $z = 14$ años $Logit = 2,18$, para $z = 82$ años, $Logit = - 1,97$ y así, diferentes riesgos según la edad

CASO CON PERFILES DE RIESGO COMPLETAMENTE ESPECIFICIDAS

Suponga que se quiere evaluar cuanto más riesgo de enfermar tiene un hombre fumador de 35 años, respecto de una mujer, fumadora de 50 años.

Entonces para:

Individuo 1, con perfil $X = 1 \quad Y = 1 \quad Z = 35$

$$\ln ODDS\ (E_1) = 1,78 + 0,168 * 1 + 1,252 * 1 - 0,061 * 35 = 1,065$$

Individuo 2, con perfil $X = 0 \quad Y = 1 \quad Z = 50$

$$\ln ODDS\ (E_2) = 1,78 + 0,168 * 0 + 1,252 * 1 - 0,061 * 50 = - 0,018$$

Si se restan los Logits

$$\ln ODDS\ (E_1) - \ln ODDS\ (E_2) =$$
$$1,78 - 1,78 + 0,168 * (1 - 0) + 1,252 * (1 - 1) - 0,061\ (35 - 50)$$

$$\ln ODDS\ (E_1) - \ln ODDS\ (E_2) = 1,081$$

Pero diferencia de logaritmos es el logaritmo de un cociente, entonces

$$\ln \frac{\text{ODDS}(E_1)}{\text{ODDS}(E_2)} = 1{,}081$$

Como: $\dfrac{\text{ODDS}(E_1)}{\text{ODDS}(E_2)} = \hat{OR}$

$$\ln \hat{OR} = 1{,}081$$

$$\therefore \quad \hat{OR} = e^{1{,}081} = 2{,}95$$

O sea un hombre fumador de 35 años, tiene 2,95 veces más riesgo de enfermar que una mujer fumadora de 50 años (cuando el riesgo en un individuo se mide con un ODDS).

Ahora expresaremos esto mismo pero con un riesgo relativo RR.

Para el Individuo 1 si:

$$\ln \text{ODDS}(E_1) = 1{,}065$$

$$\text{ODDS}(E_1) = e^{1{,}065} = 2{,}9$$

$$\text{PROB}(E_1) = \frac{2{,}9}{1+2{,}9} = 0{,}74$$

Y para el individuo 2 si:

$$\ln \text{ODDS}(E_2) = -\,0{,}018$$

$$\text{ODDS}(E_2) = e^{-0{,}018} = 0{,}98$$

$$\text{PROB}(E_2) = \frac{0{,}98}{1+0{,}98} = 0{,}495$$

Luego

$$\hat{RR} = \frac{\text{PROB}(E_1)}{\text{PROB}(E_2)} = \frac{0{,}74}{0{,}495} = 1{,}49$$

O sea un hombre fumador de 35 años, tiene 1,49 veces más riesgo de enfermar que una mujer fumadora de 50 años (cuando el riesgo en un individuo se mide con una probabilidad).
Se trata simplemente de otra forma de comparar riesgos, esta vez como un cociente de probabilidades.

Pero es necesario ahora hacer una observación importante.
Si el diseño del estudio es transversal, o de cohortes (como en este ejemplo) todos los cálculos y estimadores anteriores son válidos. Sin embargo, en .....se mencionó que en el caso de un diseño de caso y controles, los estimadores que proporciona el método de máxima verosimilitud en Regresión Logística, son válidos, salvo el de la constante del modelo.
Supongamos por un momento, que en este ejemplo, el diseño del estudio hubiese sido de casos y controles.
Con estos mismos datos, el soft de RL habría proporcionado las mismas estimaciones de los coeficientes.

$$\hat{\alpha} = 1{,}78 \qquad \hat{\beta}_1 = 0{,}168 \qquad \hat{\beta}_2 = 1{,}252 \qquad \hat{\beta}_3 = -0{,}061$$

Sin embargo ahora $\hat{\alpha} = 1{,}78$, no es un estimador válido del verdadero α poblacional.
Los β's, si lo son.

Que consecuencia trae esto?
En primer lugar los riesgos medidos por los logits para los individuo 1 y 2, en que obtuvimos 1,065 y –0,018 respectivamente, son INCORRECTOS. Y lo son, pues en su cálculo interviene la constante $\hat{\alpha} = 1{,}78$ que no es válida.
Además, como para cada individuo el ODDS (E) y la PROB(E) se calcula en base a su LOGIT(E) que es erróneo, también son no válidas estas medidas de riesgo.
En definitiva las tres medidas de riesgo son incorrectas.

Pero analicemos las medidas de comparación de riesgo.
Al ser no válidas $PROB(E_1) = 0{,}74$ y $PROB(E_2) = 0{,}495$, resulta también no válido el $\hat{RR} = 1{,}49$.
Luego bajo un diseño de casos y controles, no se puede usar el $\hat{RR}$ para comparar riesgos.

Aparentemente todos los indicadores pierden válidez.
Pero vemos que pasa con el $\hat{OR}$.
Para calcularlo, primero restamos en [3], los logits de los dos individuos. Obtuvimos como diferencia 1,081.
Pero nótese, que al restar, la constante $\hat{\alpha} = 1{,}78$ se cancela.
Esto quiere decir, que aunque $\hat{\alpha} = 1{,}78$ sea un valor erróneo, la diferencia de logits 1,081 es un valor válido, ya que depende solo de los coeficientes $\hat{\beta}_1$, $\hat{\beta}_2$ y $\hat{\beta}_3$ que sí son válidos.
Y de aquí resulta, que al ser válido 1,081, también lo es el $\hat{OR}$

$$\hat{OR} = e^{1{,}081} = 2{,}95$$

En consecuencia, para todo diseño, aún con casos y controles, el indicador $\hat{OR}$ para comparar riesgos es siempre válido.

SEPARACIÓN DE RIESGOS

El ejemplo anterior no es de los que más interesan en Regresión Logística.
Comparamos riesgos entre dos individuos que difieren en sexo, y en edad. Solo coinciden en que ambos son fumadores.

Luego el $\hat{OR}$ = 2,95 obtenido, en parte es debido a la diferencia en sexo, y en parte a la distinta edad.
Pero no sabemos en cuanto influye el sexo, y en cuanto la edad.

La intención ahora es comparar riesgos, pero evaluando la influencia de cada variable predictora por separado.

Incremento de Riesgo Debido al Sexo

Primero queremos estudiar la influencia del sexo sobre la enfermedad.
Para ello compararemos riesgos entre dos individuos de diferente sexo, pero similares en cuanto a su actitud respecto del tabaco (ambos fumadores, o ambos no fumadores), y de la misma edad.
O sea, sus perfiles de riesgo serían:

Individuo 1 X = 1 Y = y Z = z
Individuo 2 X = 0 Y = y Z = z

Notar que estos son perfiles con especificación incompleta, ya que no esta indicado si son o no fumadores, ni su edad.
Pero, eso si, estas variables no especificadas tienen idénticos valores en los dos individuos.

Calculemos riesgos:

Individuo 1, con perfil X = 1 Y = y Z = z
$\ln \text{ODDS}(E_1) = 1{,}78 + 0{,}168 * 1 + 1{,}252 * y - 0{,}061 * z$

Individuo 2 , con perfil X = 0 Y = y Z = z
$\ln \text{ODDS}(E_2) = 1{,}78 + 0{,}168 * 0 + 1{,}252 * y - 0{,}061 * z$

Notar que ambos riesgos dependen de "y" y de "z", sin embargo al restarlo

$\ln \text{ODDS}(E_1) - \ln \text{ODDS}(E_2) =$

$1{,}78 - 1{,}78 + 0{,}168 * (1-0) + 1{,}252\ (Y-Y) - 0{,}061\ (Z-Z)$

Se cancelan y:

$$\ln \frac{ODDS(E_1)}{ODDS(E_2)} = 0{,}168$$

$$\hat{OR}_x = 0{,}168$$

$$\hat{OR}_x = e^{0{,}168} = 1{,}183$$

Ahora si podemos interpretar que este riesgo es solo debido al sexo, ya que los dos individuo comparamos solo difieren en el sexo.
O sea, los varones tienen 1,183 veces más de riesgo que las mujeres, de contraer la enfermedad (a igualdad de los restantes variables)

Incremento de Riesgo Debido al Tabaco:

Si efectuamos un análisis similar para la variable Y, comparando los perfiles:

Individuo 1 con X = x Y = 1 Z = z
Individuo 2 con X = x Y = 0 Z = z

Llegamos a que:

$$\hat{OR}_Y = e^{1{,}252} = 3{,}5$$

Y este riesgo es solo debido al tabaco.
O sea, los que fuman tienen 3,5 veces más de riesgo que las que no fuman, de contraer la enfermedad (a igualdad de las restantes variables)

Incremento de Riesgo Debido a la Edad:

Para analizar la influencia de la edad, consideramos dos individuos que solo difieren en su edad. Por ejemplo uno cuya edad es Z = a, y otro de edad Z = b
Calculemos riesgos:

Individuo 1 con X = x Y = y Z = a
$\ln ODDS(E_1) = 1{,}78 + 0{,}168 * x + 1{,}252 * y - 0{,}061 * a$

Individuo 2 con X = x Y = y Z = b
$\ln ODDS(E_2) = 1{,}78 + 0{,}168 * x + 1{,}252 * y - 0{,}061 * b$

$\ln ODDS(E_1) - \ln ODDS(E_2) =$

$1{,}78 - 1{,}78 + 0{,}168 * (x-x) + 1{,}252\ (Y-Y) - 0{,}061\ (a-b)$

$$\ln \frac{ODDS(E_1)}{ODDS(E_2)} = -0{,}061 * (a-b)$$

$$\ln \hat{OR}_z = -0{,}061 * (a-b)$$

$$\hat{OR}_z = e^{-0{,}06\,1\,*\,(a-b)}$$

Notar que el incremento de riesgo depende de la diferencia de edades d = a – b, entre los individuos considerados.

$$\hat{OR}_z = e^{-0{,}061\,d}$$

Supongamos que el individuo 1 es 10 años mayor que el individuo 2, o sea d = 10.

$$\therefore \quad \hat{OR}_z = e^{-0{,}061 * 10} = 0{,}54$$

Esto quiere decir que a igualdad de las restantes variables, por cada 10 años de aumento de edad, el riesgo de contraer la enfermedad se incrementa 0,54 veces.
Como se sabe, al ser este incremento menos que 1, esto en realidad equivale a una disminución de riesgo. O sea el riesgo disminuye al 54%.
Pero esto es una particularidad de este ejemplo, ya que en general, en las enfermedades, los riesgos aumentan con la edad.
Para concluir, obsérvese que se ha estudiado separadamente el incremento de riesgo debido a cada variable.
Y para lograr esto, se comparan perfiles de riesgo que diferían en la variable de interes, y sin especificar en las restantes, pero con la condición que asuman idénticos valores.
Como medida de incremento de riesgo se utilizó el OR.

TEST DE HIPÓTESIS SOBRE LOS COEFICIENTE DEL MODELO

Ahora se describirá otra forma de evaluar la significación de los coeficientes en un modelo de RL.
Supondremos que en una investigación con 800 personas sobre una enfermedad y cinco variables de riesgo X_1, X_2, X_3, X_4, y X_5 postulamos el modelo:

$$\ln ODDS(E) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

Pero queremos ver si la variable X_5 es “importante” en el sentido de si su presencia en el modelo hace un aporte significando a la predicción de la enfermedad.

En otras palabras queremos probar la hipótesis nula:

$$H_0 : \beta_5 = 0$$

Los pasos son los siguientes:

1) ajustamos el modelo (Sin X_5):

$$\ln ODDS(E) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

Obteniendo las coeficientes estimados, y como medida del error de ajuste

$$-2 \ln L_1 = 796{,}906$$

2) ajustamos el modelo (Con X_5):

$$\ln ODDS(E) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

Obteniendo otros coeficientes estimados, y un nuevo error de ajuste

$$-2 \ln L_1 = 791{,}242$$

Notar que en general, este error de ajuste será menor que el anterior, ya que este modelo es más completo

TABLAS

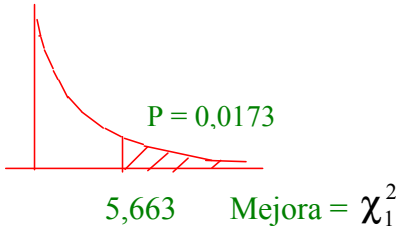
Entoces para evaluar en cuanto mejoró el ajuste debido a la inclusión de X_5, restamos estas cantidades

$$(-2 \ln L_1) - (-2 \ln L_2) = 796{,}906 - 791{,}242$$

$$(-2 \ln L_1 - \ln L_2) = -2 \ln \left[\frac{L_1}{L_2}\right] = 5{,}663$$

Pero, esta mejora es importante? O la podemos considerar despreciable?.
Para responder esto, necesitamos algún valor para compararar.
Usamos el siguiente resultado matemático que dice:

"Si el verdadero coeficiente β_5 es nulo, o sea si $\beta_5 = 0$, al ajustar el segundo modelo con X_5, la mejora se distribuye aproximadamente como un Chi Cuadrado con 1 grado de libertad"

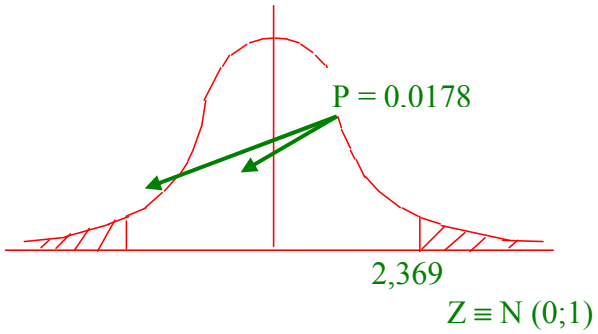


Se concluye entonces que la mejora es importante, y con un nivel de significación de P = 0,0173 se rechaza la hipótesis $\beta_5 = 0$.

Es importante mencionar que para que la mejora se distribuya aproximadamente como una χ^2, es necesario que la muestra del estudio sea grande en relación a cantidad de variables del modelo.
Y en este caso eso se cumple ya que n = 800 es mucho mayor que 5 (cantidad de variables)

Como ya fue mencionado antes, otra manera de testear si $\beta_5 = 0$, es utilizado el desviante normal (ver tabla de coeficientes del 2º modelo.

$$Z = \frac{\beta_5}{SE} = \frac{0,1295}{0,0546} = 2,369$$



Sin embargo, aunque en este ejemplo las **p** obtenidas por ambos métodos fueron casi iguales (0,0173 y 0,0178) en muestras menores, hay algunos estudios que demuestran que el test de razón de verosimilitud es superior (Jennings 1986).

CASO DE TRES COEFICIENTES

Siguiendo con el ejemplo anterior suponga que ahora queremos averiguar si en conjunto, las variables X_3, X_4, y X_5 hacen algún aporte a la predicción de la enfermedad.

Concretamente queremos evaluar la hipótesis

$$H_0: \beta_3 = 0, \quad \beta_4 = 0, \quad \beta_5 = 0,$$

Procediendo como en el ejemplo anterior, primero ajustamos el modelo

$$\ln ODDS(E) = \alpha + \beta_1 X_1 + \beta_2 X_2$$

obteniendo los coeficientes y

$$-2 \ln L_1 = 1043,318$$

Luego ampliado con X_3, X_4, y X_5

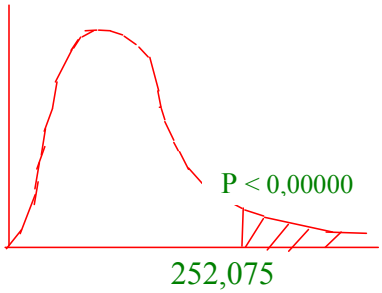
$$\ln ODDS(E) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

Obtenemos otros coeficientes y $-2 \ln L_2 = 791,242$

Ahora la mejora es:

Mejora = 1043,318 – 791,242 = 252,075

Pero ahora, como son tres las variables agregadas, esta mejora, en el caso de ser cierta la hipótesis nula H_0 se distribuye como una χ^2 con 3 grados de libertad



Mejora ≡ χ^2_3

Como P < 0,0000 se rechaza con este nivel de significación altísimo en este ejemplo la hipótesis de que los tres coeficientes son nulas ($\beta_3 = \beta_4 = \beta_5 = 0$)

La interpretación práctica de este rechazo, es que alguno (todos) de los tres coeficientes es ≠ de 0.
Pero no sabemos cual, o cuales.
Para averiguarlo, deberíamos hacer tests por separado como en el punto anterior

CASO GENERAL

Integrado todo lo anterior, resulta que si se tiene el modelo

$$\ln ODDS\ (E) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots + \beta_k X_k$$

y queremos probar la hipótesis de que los últimas r coeficientes son nulos:

$$H_0: \beta_{r+1} = 0 \quad \beta_{r+2} = 0 \quad \beta_k = 0$$

r coeficientes

Entonces primero se ajusta un modelo reducido sin las variables de estos coeficientes

$$\ln ODDS\ (E) = \alpha + \beta_1 X_1 + \ldots\ldots + \beta_r X_r$$

Que tiene (k-r) + 1 coeficientes (acordarse de α)
Obteniendo:

$$-2 \ln L_1$$

Luego se ajusta el modelo ampliado con las r variables

$$\ln ODDS\ (E) = \alpha + \beta_1 X_1 + \beta_r X_r + \beta_{r+1} X_{r+1} + \beta_k X_k$$

Que tiene k+1 coeficientes (los de la k variables y el de la constante)
Obteniendo:

$$-2 \ln L_1$$

A continuación se calcula la "mejora" en el ajuste

$$\text{Mejora} = (-2 \ln L_1) - (-2 \ln L_2)$$

Que en el caso de ser cierta H_0, se distribuye como una χ^2, cuyos grados de libertad corresponden a la diferencia entre la cantidad de coeficientes de los dos modelos, o sea:

$$\nu = \underbrace{(k+1)}_{\text{modelo ampliado}} - \underbrace{(k-r+1)}_{\text{modelo reducido}} = r$$

Que coincide con la cantidad de coeficiente a testear en H_0 (o sea r)

Si operamos matemáticamente sobre la expresión de la mejora en el ajuste de [1] resulta:

$$\text{Mejora} = (-2 \ln L_1) - (-2 \ln L_2)$$

$$= -2 (\ln L_1 - 2 \ln L_2)$$

$$= -2 \ln \left[\frac{L_1}{L_2} \right]$$

Nos queda la mejora expresada como cociente o razón de las verosimilitudes
L_1 y L_2

De aquí que a este test se le llame test de razón de verosimilitudes

En definitiva resulta

$$\text{Mejora} = -2 \ln \frac{L_1}{L_2} \rightarrow \chi^2_r$$

Donde r es número de coeficientes a testear.

NOTA SOBRE EL OUTPUT DE UN SOFT DE RL

Cuando se ajusta un modelo como:

$$\ln \text{ODDS}(E) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Usualmente los soft proporcionarán además de la tabla con los coeficientes estimados

	B	S.E	Wald	Sig.
X_1	0,080	0,161	0,250	0,617
X_2	-0,076	0,150	0,259	0,611
X_3	0,397	0,155	6,555	0,010
	-0,204	0,139	2,151	0,142

La verosimilitud del modelo, expresado como:

$- 2 \ln L$, o sea

$- 2 \ln L = 1101,352$

y además un test global para los coeficientes, que en este caso es:

Test Global

Cai-Square	df	Sig.
6,963	3	0,073

Donde: df son los grados de libertad

Este test prueba la hipótesis si las coeficientes de las tres variables son nulas, o sea

$$H_0: \beta_1 = 0 \ , \quad \beta_2 = 0 \ , \ \beta_3 = 0$$

Sirve para evaluar, si en conjunto, las tres variables del modulo propuesto hacen algún aporte significativo a la predicción de la enfermedad.
En este ejemplo, como p = 0,073>0,05, se decide no rechazar la hipótesis nula, al menos para un nivel de significativo del 5%

Pero como realiza esta prueba el soft?
Primero, ajusta el modelo reducido, sin las variables (solo con la constante

$$\text{Ln ODDS (E)} = \alpha$$

Obteniendo $- 2 \ln L_1 = 1108,315$

Luego ajusta el modelo ampliado, con las tres variables y la constante

$$\text{Ln ODDS (E)} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Obteniendo como se vio

$$- 2 \ln L_2 = 1101,352$$

Luego la Mejora

$$\text{Mejora} = 1108,315 - 1101,352 = 6,963$$

Se distribuye como una χ_3^2 con tres grados de libertad, y son estas cantidades, junto a la significación, las que aparecen en la tabla del test global.