Pontificia Universidad JAVERIANA Bogotá

MAESTRÍA EN EPIDEMIOLOGÍA CLÍNICA

BIOESTADÍSTICA AVANZADA

MÓDULO II

Semana 10

Regresión Logística

Material elaborado por:

Nelcy Rodriguez Malagón
Bioestadística, M.P.H
Profesora Titular
Departamento de Epidemiología Clínica y Bioestadística
Facultad de Medicina - Pontificia Universidad Javeriana

Kevin Maldonado-Cañón
Médico Epidemiólogo, MD. MSc.
Estudiante - Doctorado en Epidemiología Clínica
Departamento de Epidemiología Clínica y Bioestadística
Facultad de Medicina - Pontificia Universidad Javeriana

Modelo de Regresión Logística para evaluar la asociación entre un factor de exposición y un desenlace considerando además la presencia de potencial interacción o confusión

Introducción

En el estudio de la relación entre la ocurrencia de unos factores o variables y un desenlace en salud y especialmente en clínica, con no poca frecuencia aparece la pregunta si existe alguna relación entre uno de esos factores identificado como de interés y la ocurrencia de un desenlace o si ya ocurrido un desenlace, puede haber alguna diferencia entre quienes se consideraban expuestos al factor de interés y quienes no estaban expuestos al mismo. En este caso, ya no se trata de predecir un evento con base en un conjunto de factores o variables independientes como se revisó previamente.

Podemos preguntar, por ejemplo, si el sobrepeso puede estar asociado a la ocurrencia de diabetes mellitus tipo II cuando además consideramos factores como la edad, el hábito de fumar, el sexo, el sedentarismo y la dieta. En este caso, el factor de interés o de exposición es el sobrepeso, el desenlace es la ocurrencia de la diabetes mellitus y los otros factores mencionados, constituyen covariables cuyo efecto debería ser controlado en el estudio de la relación de interés.

En el contexto anterior y dado que no se tienen solamente dos variables de interés, parece lógico recurrir al uso del modelamiento estadístico para el estudio del problema. Dado que la variable desenlace es cualitativa y especialmente dicótoma (aunque también puede ser policótoma u ordinal), el modelo de regresión logística es una herramienta que se considera y que se ha utilizado ampliamente.

Nota: aquí revisaremos el caso más simple que corresponde a tener un factor de exposición de carácter dicótomo y un desenlace igualmente dicótomo. Sin embargo, se puede tener más de un factor de exposición en algunas situaciones. Esos factores de exposición pueden ser de tipo policótomo u ordinales. También el desenlace puede ser tanto policótomo como ordinal.

En esta situación, volvemos a considerar el modelo:

$$f(y) = \frac{1}{1 + e^{-y}}\ ,\ -\infty < y < +\infty$$

Donde $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$

$$f(y) = Pr(D = 1 | x_2, x_2, \ldots x_k) = P(x)$$

Entonces: $P(x) = \frac{1}{1+e^{-(\alpha+\sum \beta_j x_j)}}$

Expresión que al ser linealizada corresponde a: $logit[P(x)] = \alpha + \sum \beta_j x_j$

En este caso, usualmente la primera variable independiente es el factor de exposición cuya relación con el desenlace es el interés primario y el resto de variables xj corresponden a los potenciales factores de confusión.

Cómo ¿entonces podemos usar el modelo de regresión logística para estudiar la asociación entre la exposición y un desenlace?

En este caso, podemos considerar una estrategia que implica los siguientes pasos:

1. Definir o especificar las variables que se consideran en la pregunta de investigación y que constituyen el desenlace, el o los factores de estudio (factor o factores de exposición), las potenciales variables de confusión y las potenciales interacciones
2. Evaluar la presencia de posible interacción o interacciones
3. Evaluar la presencia de confusión

Definición de variables a incluir en el modelo

En este punto, se deben especificar todas las variables a ser consideradas de interés en el problema y susceptibles de ser incluidas en un modelo inicial.

- Es importante anotar que en cuanto a los factores de riesgo se deberían incluir aquellos que son conocidos pro literatura o que se sospecha pueden serlo para disminuir la posibilidad de encontrar asociaciones espúreas.
- Algunas veces y cuando se tienen muchos factores de riesgo, se tiende a seleccionar para a incluir al modelo, aquellos que resultan significativos en análisis iniciales exploratorios. Sin embargo, no es una recomendación buena hacer esto olvidando que la significancia clínica es primero. En este proceso de selección de variables, se debe tener en cuenta el tamaño de la muestra revisado en temas anteriores para disminuir la posibilidad de posibles sobreajustes de modelo.
- La exploración de una potencial presencia de colinealidad entre las variables independientes, debe ser tenida en cuenta antes de correr un modelo.
- Es importante además determinar de antemano aquellas potenciales interacciones que pueden definirse entre la exposición y potenciales factores de confusión antes de correr un modelo. Aquí se debe recordar que ellas se definen especialmente desde el punto de vista clínico y que cuando se consideran, además se consideran más parámetros a ser estimados.
- Existe un principio jerárquico cuando se definen interacciones en un modelo: Si una interacción se incluye en el modelo, entonces los términos que se incluyen en ella, deben ser mantenidos en éste. Por ejemplo:
    - Si E: es la exposición, $X3$ es un factor independiente y potencial variable de confusión y se define la interacción $EX3$ entonces tanto la exposición como $X3$ deben estar en el modelo y también la interacción en caso de que ella resultase significativa al correr dicho modelo. Igual puede ocurrir con interacciones de mayor orden.

Evaluación de la presencia de interacción Una metodología recomendada en el proceso de llegar a un modelo que refleje mejor la relación que se estudia entre exposición y desenlace es la estrategia de eliminación de términos del modelo en forma jerárquica. Así:

1. Iniciar con el modelo completo
2. Eliminar términos de interacción no importantes o significativos en el modelo
3. Eliminar potenciales factores de confusión no importantes en el modelo.

¡Siempre se debe evaluar primero la interacción que la confusión!

Si existe una interacción significativa entre la exposición y un factor potencial de confusión, ya no debe evaluarse si esa es una variable de confusión.

Para evaluar los términos de interacción, se pueden efectuar pruebas globales a fin de probar el efecto de todos o algunos de esos términos de interacción entre la exposición y las variables de confusión. Si alguno de estos términos es significativo, debe permanecer en el modelo tanto la interacción como el potencial factor de confusión involucrado en ella.

Evaluación de la presencia de confusión

En el caso de la evaluación de presencia de confusión pueden ocurrir dos situaciones:

1. Si alguna interacción fue significativa: Inicie con el modelo que contiene la exposición, todos los potenciales factores de confusión y el o los términos de interacción encontrados significativos previamente. No olvide que no puede excluir los potenciales factores de confusión que son parte de las interacciones que fueron significativas. Si es posible, elimine del modelo otros factores que pueden ser de confusión a fin de incrementar precisión, manteniendo validez.

2. Si ninguna interacción definida, fue significativa: Inicie con el modelo que contiene tanto la exposición como los potenciales factores de confusión definidos. Luego remueva uno o más de ellos del modelo y evalúe en cada paso si hay un cambio en el riesgo relativo indirecto (OR).

3. Si no hay cambio en el OR, puede dejar o no los potenciales factores de confusión. La decisión depende de la precisión y de razones clínicas. Si la no precisión mejora, pero clínicamente es un término importante, quizá se debe mantener el término en el modelo.

Ejemplo práctico

Para ejemplificar todo el proceso, usaremos la base de datos `icu`, que hemos venido trabajando en sesiones pasadas. En este caso, estamos interesados en determinar si hay una asociación entre el estado vital del paciente (`sta`) y su servicio de procedencia (`ser`), teniendo en cuenta además las variables infección, edad (`age`) y sexo (`gender`). No olvidemos que es muy importante recordar la forma de codificación de las variables antes de iniciar el proceso de análisis.

Iniciamos con el modelo completo y obtenemos los coeficientes de regresión de este modelo:

```
## Recuerden siempre al inicio de cada ejercicio cargar la base de datos
base_icu <- aplore3::icu
# Asignamos "Surgical" como nuestra categoría de referencia
base_icu$ser <- relevel(base_icu$ser, ref = "Surgical")

modelo <- glm(
  sta ~ ser + inf + age + gender,
  data = base_icu,
  family = binomial
)

summary(modelo)
```

```
##
## Call:
## glm(formula = sta ~ ser + inf + age + gender, family = binomial,
##     data = base_icu)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.84034    0.80187  -4.789 1.67e-06 ***
## serMedical    0.87826    0.38919   2.257    0.024 *
## infYes        0.61090    0.38252   1.597    0.110
## age           0.02787    0.01134   2.457    0.014 *
## genderFemale -0.01936    0.37974  -0.051    0.959
## ---
## 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: 182.07  on 195  degrees of freedom
## AIC: 192.07
##
## Number of Fisher Scoring iterations: 5
```

Con este comando se obtienen los coeficientes de regresión y las estadísticas para su evaluación.

Al elevar al exponente el valor del coeficiente para la variable servicio (`ser`), se obtiene el riesgo relativo indirecto ajustado por las variables:

```
## Forma manual de exponenciar los coeficientes del modelo
exp(coef(modelo))
##   (Intercept)    serMedical         infYes           age genderFemale
##    0.02148635    2.40670121    1.84208002    1.02825786    0.98083052
## Uso la función tbl_regression para exponenciar los coeficientes del modelo
gtsummary::bold_p(gtsummary::tbl_regression(modelo, exponentiate = TRUE, pvalue_fun =
~ gtsummary::style_pvalue(.x, digits = 2)))
```

Characteristic	OR	95% CI	p-value
ser			
Surgical	—	—	
Medical	2.41	1.14, 5.27	**0.024**
inf			
No	—	—	
Yes	1.84	0.87, 3.94	0.11
age	1.03	1.01, 1.05	**0.014**
gender			
Male	—	—	
Female	0.98	0.46, 2.05	0.96

Abbreviations: CI = Confidence Interval, OR = Odds Ratio

Ahora bien, note que el modelo más simple en el estudio de la asociación se puede procesar mediante el código:

```
modelo <- glm(
  sta ~ ser,
  data = base_icu,
  family = binomial
)


summary(modelo)
```

```
##
## Call:
## glm(formula = sta ~ ser, family = binomial, data = base_icu)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.8935     0.2867  -6.605 3.97e-11 ***
## serMedical    0.9469     0.3682   2.572   0.0101 *
## ---
## 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: 193.24  on 198  degrees of freedom
## AIC: 197.24
##
## Number of Fisher Scoring iterations: 4
```

```
gtsummary::bold_p(gtsummary::tbl_regression(modelo, exponentiate = TRUE, pvalue_fun =
~ gtsummary::style_pvalue(.x, digits = 2)))
```

Characteristic	OR	95% CI	p-value
ser			
Surgical	—	—	
Medical	2.58	1.27, 5.43	**0.010**

Abbreviations: CI = Confidence Interval, OR = Odds Ratio

Observe que el valor obtenido aquí es el mismo que el encontrado al hacer el análisis crudo de los datos. Esto es al obtener mediante una tabla de contingencia, el valor del OR.

Evaluación de la Interacción

Para evaluar si algunos de los factores considerados pueden ser modificadores de efecto (presentes en una interacción), o potenciales factores de confusión, se define las posibles interacciones y se incluyen en el modelo.

Recordemos que una interacción ocurre entre la exposición y un potencial factor de confusión.

Para el caso, la interacción de interés es la que se presume entre la variable servicio y la variable edad. Para ello se define la variable `ser*age`. También pueden definirse otras interacciones simples o complejas.

Este es el modelo completo o saturado (con la interacción):

```
modelo_1 <- glm(
  sta ~ ser*age + inf + gender,
  data = base_icu,
  family = binomial
)

summary(modelo_1)
```

```
##
## Call:
## glm(formula = sta ~ ser * age + inf + gender, family = binomial,
##     data = base_icu)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.314258   0.886120  -2.612  0.00901 **
## serMedical     -2.339996   1.485317  -1.575  0.11516
## age             0.003136   0.014241   0.220  0.82570
## infYes          0.554242   0.391492   1.416  0.15686
## genderFemale    0.098769   0.393105   0.251  0.80162
## serMedical:age  0.051233   0.022939   2.233  0.02552 *
## ---
## 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: 176.91  on 194  degrees of freedom
## AIC: 188.91
##
## Number of Fisher Scoring iterations: 5
```

```
gtsummary::bold_p(gtsummary::tbl_regression(modelo_1, exponentiate = TRUE, pvalue_fun
= ~ gtsummary::style_pvalue(.x, digits = 2)))
```

Characteristic	OR	95% CI	p-value
ser			
Surgical	—	—	
Medical	0.10	0.00, 1.71	0.12
age	1.00	0.98, 1.03	0.83
inf			
No	—	—	
Yes	1.74	0.81, 3.78	0.16
gender			
Male	—	—	
Female	1.10	0.50, 2.37	0.80
ser * age			
Medical * age	1.05	1.01, 1.10	**0.026**

Abbreviations: CI = Confidence Interval, OR = Odds Ratio

Este constituye el modelo simple (sin la interacción):

```
modelo_2 <- glm(
  sta ~ ser + inf + age + gender,
  data = base_icu,
  family = binomial
)

summary(modelo_2)
```

```
##
## Call:
## glm(formula = sta ~ ser + inf + age + gender, family = binomial,
##     data = base_icu)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.84034    0.80187  -4.789 1.67e-06 ***
## serMedical    0.87826    0.38919   2.257    0.024 *
## infYes        0.61090    0.38252   1.597    0.110
## age           0.02787    0.01134   2.457    0.014 *
## genderFemale -0.01936    0.37974  -0.051    0.959
## ---
## 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: 182.07  on 195  degrees of freedom
## AIC: 192.07
##
## Number of Fisher Scoring iterations: 5
```

```
gtsummary::bold_p(gtsummary::tbl_regression(modelo_2, exponentiate = TRUE, pvalue_fun
= ~ gtsummary::style_pvalue(.x, digits = 2)))
```

Characteristic	OR	95% CI	p-value
ser			
Surgical	—	—	
Medical	2.41	1.14, 5.27	**0.024**
inf			
No	—	—	
Yes	1.84	0.87, 3.94	0.11
age	1.03	1.01, 1.05	**0.014**
gender			
Male	—	—	
Female	0.98	0.46, 2.05	0.96

Abbreviations: CI = Confidence Interval, OR = Odds Ratio

Para evaluar si la interacción es significativa, utilizamos el test de razón de máxima verosimilitud (*likelihood ratio test*). La estadística de prueba está dada por la siguiente expresión:

$$LR = 2ln(L\,modelo\,restringido - L\,modelo\,completo)$$

Donde *L* es la función máximo verosímil estimada tanto para el modelo completo (con todas las variables incluidas) como para el modelo simple (sin incluir la variable que se quiere probar: en este caso, la interacción).

La estadística LR sigue una distribución chi-cuadrado con grados de libertad igual al número de parámetros del modelo completo menos el número de parámetros del modelo simple.

```
# Comparación con LR test
anova(modelo_1, modelo_2, test = "LRT")
## Analysis of Deviance Table
##
## Model 1: sta ~ ser * age + inf + gender
## Model 2: sta ~ ser + inf + age + gender
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       194     176.91
## 2       195     182.07 -1  -5.1585  0.02313 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Dado que la prueba fue significativa, la interacción debe ser entonces considerada como un término importante en el modelo y en la interpretación del riesgo relativo. Note que entonces **el riesgo que debería interpretarse es el de la interacción** y no necesariamente el del potencial factor de confusión (que en este caso es la edad).

Note que el término de interacción puede evaluarse mediante la prueba de Wald, cuya significancia está reportada automáticamente o mediante el método de Máxima Verosimilitud, utilizando la estadística L (Log Likelihood). Una vez procesados los resultados, los valores de estas dos estadísticas son aproximadamente iguales.

Se podría continuar con el proceso de eliminación de otros factores potenciales de confusión como la variable infección y la variable sexo. Ello dependerá de la ganancia en precisión como se mencionó y de la importancia clínica.

Evaluación de la presencia de variables de confusión

Supongamos que interesa evaluar si la variable infección es un factor de confusión.

Del modelo que incluye la infección, excluimos esta variable y comparamos con el siguiente:

```
modelo <- glm(
  sta ~ ser + age + gender,
  data = base_icu,
  family = binomial
)

summary(modelo)
```

```
##
## Call:
## glm(formula = sta ~ ser + age + gender, family = binomial, data = base_icu)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.76291    0.79776  -4.717  2.4e-06 ***
## serMedical    1.02189    0.37860   2.699  0.00695 **
## age           0.03029    0.01121   2.702  0.00689 **
## genderFemale -0.03524    0.37690  -0.094  0.92550
## ---
## 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: 184.64  on 196  degrees of freedom
## AIC: 192.64
##
## Number of Fisher Scoring iterations: 5
```

```
gtsummary::bold_p(gtsummary::tbl_regression(modelo, exponentiate = TRUE, pvalue_fun =
~ gtsummary::style_pvalue(.x, digits = 2)))
```

Characteristic	OR	95% CI	p-value
ser			
Surgical	—	—	
Medical	2.78	1.34, 5.98	**0.007**
age	1.03	1.01, 1.06	**0.007**
gender			
Male	—	—	
Female	0.97	0.45, 2.01	0.93

Abbreviations: CI = Confidence Interval, OR = Odds Ratio

De los resultados podemos decir que efectivamente **la variable infección es un factor de confusión**.

Evaluemos ahora la variable edad. Comparamos de nuevo los dos modelos: con y sin edad.

```
modelo <- glm(
  sta ~ ser + inf + gender,
  data = base_icu,
  family = binomial
)

summary(modelo)
```

```
##
## Call:
## glm(formula = sta ~ ser + inf + gender, family = binomial, data = base_icu)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.19618    0.35922  -6.114 9.73e-10 ***
## serMedical    0.79323    0.37857   2.095   0.0361 *
## infYes        0.75868    0.37168   2.041   0.0412 *
## genderFemale  0.03952    0.37286   0.106   0.9156
## ---
## 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: 189.01  on 196  degrees of freedom
## AIC: 197.01
##
## Number of Fisher Scoring iterations: 4
```

```
gtsummary::bold_p(gtsummary::tbl_regression(modelo, exponentiate = TRUE, pvalue_fun =
~ gtsummary::style_pvalue(.x, digits = 2)))
```

Characteristic	OR	95% CI	p-value
ser			
Surgical	—	—	
Medical	2.21	1.06, 4.73	**0.036**
inf			
No	—	—	
Yes	2.14	1.04, 4.48	**0.041**
gender			
Male	—	—	
Female	1.04	0.49, 2.15	0.92

Abbreviations: CI = Confidence Interval, OR = Odds Ratio

Decisión Clínica: ¿Qué modelo seleccionamos para evaluar la relación entre el estado vital y el servicio de procedencia de los pacientes?

Nota: Desde el punto de vista matemático: Es bueno mantener todos los factores de confusión, pero puede reducirse eventualmente la precisión en la estimación del riesgo indirecto.

Lecturas complementarias:

- Kleinbaum DG, Klein M. Logistic regression: a self-learning text. 3rd ed. New York: Springer; 2010. Capítulos 6 y 7