Pontificia Universidad JAVERIANA Bogotá

MAESTRÍA EN EPIDEMIOLOGÍA CLÍNICA

BIOESTADÍSTICA AVANZADA

MÓDULO II

Semana 7

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

Introducción

Con bastante frecuencia en investigación clínica, una pregunta de interés por responder es cómo una serie de factores puede estar **asociada** con la ocurrencia de un evento o resultado. Por ejemplo: *¿Qué condiciones pueden estar asociadas a la ocurrencia de un infarto agudo de miocardio?* o *¿Cuáles pueden estar asociadas con la presentación de una pancreatitis aguda?* También puede ocurrir que la pregunta de interés sea cuáles de esos factores pueden **predecir** la ocurrencia de alguna de estas patologías.

Para encontrar una respuesta a cualquiera de estos dos problemas, el modelamiento estadístico, específicamente el **modelo de regresión logística**, constituye una muy buena herramienta de análisis al considerar como variable de desenlace un evento de interés cualitativo cuya observación en cada individuo produce un **dato dicótomo**.

Supuestos y planteamiento del modelo

Función logística

El escenario estadístico es claro que:

- Tenemos un desenlace binario (p.ej. D = 1 enfermo, D = 0 no enfermo)
- Lo que queremos modelar es una probabilidad, es decir, un número entre 0 y 1

La regresión logística se basa en la función logística, definida como:

$$f(z) = \frac{1}{1 + e^{-z}}$$

donde:

- e es la base del logaritmo natural
- z puede tomar cualquier valor real desde − ∞ hasta + ∞

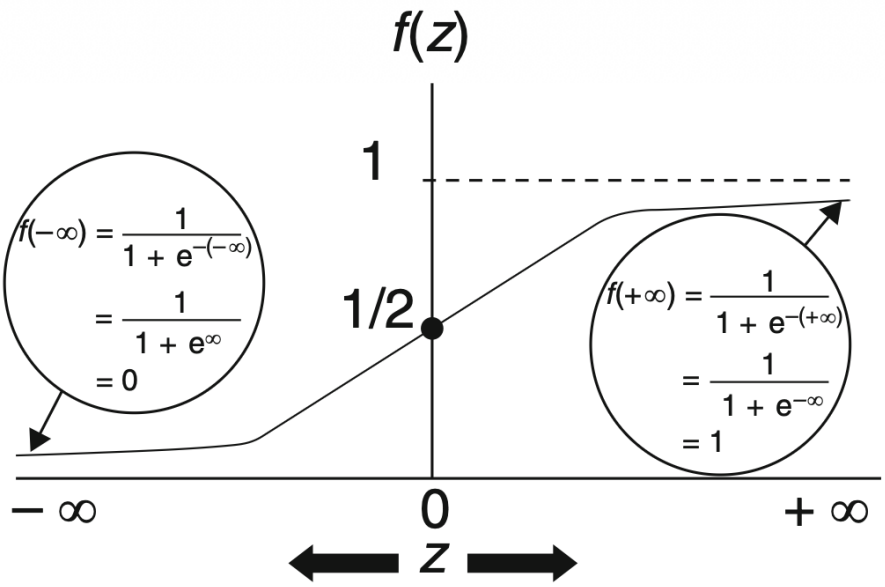
La función logística siempre produce valores entre 0 y 1, es decir, valores compatibles con una probabilidad.

En epidemiología, una probabilidad representa:

$P(D=1)$
es decir, el riesgo de desarrollar el evento de interés (D).

La forma en S de la función logística

La gráfica de la función logística tiene forma de S alargada, como se puede ver a continuación:



Kleinbaum, D. G., & Klein, M. (2010). Logistic Regression: A Self-Learning Text (3rd ed.). Springer
Cuando aumentamos z:

1. Para valores muy bajos de z, la probabilidad se mantiene cercana a 0
2. Luego hay un rango donde la probabilidad aumenta rápidamente
3. Finalmente, se aplana cerca de 1

Si interpretamos z como una combinación de una o más variables independientes entonces:

- Para exposiciones bajas → poco efecto
- Al alcanzar cierto "umbral" → el riesgo aumenta rápidamente
- Para exposiciones muy altas → el riesgo se aproxima a 1 y se estabiliza

Este comportamiento es muy compatible con muchos fenómenos biológicos. Por eso la forma en S resulta conceptualmente útil.

De la función logística al modelo logístico

En un modelo de regresión, necesitamos expresar z como función de una o más variables explicativas:

$$z = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Aquí:

- $X_1, X_2, \ldots, X_k$ son variables independientes
- α es el intercepto
- β_i son parámetros desconocidos
- z es un índice lineal que combina los factores de riesgo

En esencia, entonces, z es un índice que combina las Xs.

De esta manera, podemos sustituir este z en la función logística:

$$z = \underbrace{\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k}$$

$$f(z) = \frac{1}{1 + e^{-z}}$$

$$= \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}}$$

Kleinbaum, D. G., & Klein, M. (2010). Logistic Regression: A Self-Learning Text (3rd ed.). Springer

Este, finalmente, es el modelo logístico.

Los parámetros representan:

- α: log-odds cuando todas las X son 0
- β_i: cambio en el log-odds por unidad de cambio en X_i, ajustado por las demás variables

Logit aplicado al modelo logístico

Si partimos del modelo logístico:

$$P(X) = \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}}$$

Podemos demostrar algebraicamente que:

$$\frac{P(X)}{1-P(X)} = e^{\alpha} + \sum \beta_i X_i$$

Y si ahora tomamos logaritmo natural:

$$ln(\frac{P(X)}{1-P(X)}) = \alpha + \sum \beta i Xi$$

Tenemos que:

$$logit[P(X)] = \alpha + \sum \beta_i X_i$$

- El modelo original es no lineal en probabilidad
- Pero es lineal en el log odds

La regresión logística no modela la probabilidad directamente. **Modela el logaritmo del odds.**

Por eso muchas veces vemos el modelo escrito como:

$$logit(P) = \alpha + \beta_1 X_1 + \beta_2 X_2 + ...$$

que es la forma lineal de expresar el modelo.

Al hacer uso del modelo de regresión logística, se deben cumplir los siguientes supuestos:

Supuestos

Una de las ventajas al usar el modelo de regresión logística, es que, a diferencia de un modelo de regresión lineal simple o múltiple, no se requiere cumplir con el supuesto de normalidad de las variables independientes ni del desenlace. De otro lado, una diferencia sustancial es que en un modelo de regresión lineal sea simple o múltiple, la variable o el evento de interés debe ser numérica sea de tipo discreto o continuo en tanto que, en un modelo de regresión logística, la variable de desenlace o interés es cualitativa y puede ser dicótoma o no. Tanto en el caso en que se estudie la asociación, como en el que se explore la predicción de un evento especialmente dicótomo, el modelo de regresión logística debe cumplir con los siguientes supuestos.

1. La variable dependiente debe ser binaria, puede ser dicótoma o policótoma o también medida en una escala ordinal.
2. Independencia de las observaciones: Lo cual significa que el resultado para un individuo, no depende del desenlace en otro.
3. Muestra suficiente: Especialmente para el caso en que se busque la predicción de un fenómeno, el tamaño de muestra debe ser suficientemente grande.
4. Ausencia de multicolinealidad entre las variables independientes: Las variables explicativas no requieren cumplir el supuesto de normalidad ni tener una relación lineal directa con el desenlace. Sin embargo, idealmente no deben presentar alta colinealidad entre sí, ya que la redundancia de información puede generar inestabilidad en los coeficientes, ampliar los errores estándar y dificultar la interpretación.
5. Linealidad del logit: Se espera que haya una relación lineal entre las variables independientes continuas y el logaritmo (logit) de la probabilidad de ocurrencia del desenlace o variable dependiente.
6. Observaciones influyentes: La presentación de valores que correspondan a valores influyentes o extremos, pueden resultar en un gran impacto sobre la estimación de los coeficientes.

Estimación de coeficientes e interpretación

Ajustar un modelo de regresión logística significa usar los datos observados para estimar los parámetros desconocidos α y β_i.
Una vez estimados α y los β_i, podemos:

1. Tomar un individuo con valores conocidos de X
2. Sustituirlos en la ecuación
3. Obtener su probabilidad estimada de tener el evento

Para efecto de estimación de los parámetros desconocidos α y β_i, se hace uso del método de *máxima verosimilitud*.

Método de máxima verosimilitud

Lo explicaremos con un ejemplo corto:

Supongamos que tenemos una muestra.

Para cada individuo i:

- Observamos X_i (un valor que toma una variable independiente para el individuo i)
- Observamos D_i (un valor de 0 o 1 que toma el desenlace para el individuo i)

El modelo nos da una probabilidad estimada:

$$P_i = P(D_i = 1 \mid X_i)$$

Ahora pensemos:

- Si el individuo desarrolló la enfermedad $(D = 1)$, la probabilidad de observar ese resultado es P_i.
- Si no la desarrolló $(D = 0)$, la probabilidad de observar ese resultado es $1 - P_i$.

Podemos escribir esto como:

$$P_i^{D_i}(1 - P_i)^{1-D_i}$$

Función de verosimilitud

Asumiendo independencia entre individuos (supuesto clave), la probabilidad conjunta de observar todo el conjunto de datos es:

$$L(\alpha, \beta) = \prod_{i=1}^{n} P_i^{D_i}(1 - P_i)^{1-D_i}$$

Esta es la función de verosimilitud.

Pero cuidado. Esta función no expresa la probabilidad de los parámetros, sino **la probabilidad de los datos dado un conjunto de parámetros**.

El método de máxima verosimilitud busca los valores de:

$$\alpha, \beta 1, \ldots, \beta k$$

que maximizan esa función de verosimilitud.

En otras palabras, encuentra los parámetros que hacen que los datos observados sean lo más "probables" bajo el modelo.

Estos procesos no se hacen manualmente. En cambio, los softwares o paquetes estadísticos usan algoritmos iterativos para tomar el logaritmo de la verosimilitud (log-likelihood), derivar ecuaciones y resolverlas.

¿Por qué no usamos mínimos cuadrados como en la regresión lineal?

En regresión lineal:

- El desenlace es continuo
- Los errores siguen una distribución normal
- Responde a la pregunta: ¿Qué línea reduce mejor los errores?

Pero en la regresión logística:

- El desenlace es binario
- La distribución es binomial
- La varianza depende de la media: $\mathrm{Var}(D_i)=P_i(1-P_i)$
- Queremos responder la pregunta: ¿Qué parámetros hacen que estos datos sean los más probables?
- No buscamos minimizar la distancia vertical entre lo observado y lo predicho, sino maximizamos coherencia probabilística de haber observado exactamente esos datos

Características de los estimadores que se obtienen con el método de máxima verosimilitud

Bajo condiciones regulares los estimadores:

- Son consistentes
- Son asintóticamente normales
- Permiten construir intervalos de confianza
- Permiten pruebas de hipótesis (Wald, razón de verosimilitud, score)

Interpretación de los coeficientes

Interpretación del intercepto: α

Si todos los $X_i = 0$:

$$logit(P) = \alpha$$

Entonces:

$$\alpha = log\ odds\ cuando\ todos\ los\ X\ =\ 0$$

Dado que no existe un individuo con todos los X = cero (ej: edad = 0).

Es mejor interpretar α como:

- El logaritmo del odds basal (background odds).

Es decir, el odds que resultaría en un modelo sin predictores.

Interpretación de los coeficientes: ß

Tomemos el modelo:

$$logit(P) = \alpha + \beta iX$$

Entonces:

βi = Cambio en el log odds cuando X (variable independiente) toma un valor dado, manteniendo el resto de variables constantes.

Calculo e interpretación de la medida de asociación (Exp(beta) = OR)

Sabemos que:

$$log\ odds = \alpha + \sum \beta iXi$$

Si β representa un cambio en log odds, entonces al exponenciar:

$$e^{\beta}{}_{L}$$

obtenemos el cambio multiplicativo en el odds.

En otras palabras:

- β está en escala log
- e^{β} está en escala de odds ratio (OR) (siendo el OR una medida epidemiológicamente interpretable)

Pruebas de hipótesis sobre coeficientes y las medidas de asociación

En un modelo de regresión logística, las pruebas de hipótesis sobre los coeficientes pueden formularse tanto en términos del parámetro β (en escala logit) como en términos del Odds Ratio (OR), que es su transformación exponencial. Conceptualmente, estamos evaluando si existe evidencia estadística de un efecto entre una variable independiente X_i y el evento de interés.

Hipótesis en términos de βi

Para cada coeficiente β_i, se plantean:

- Hipótesis nula (H_0) $H0: \beta_i = 0$
- Hipótesis alternativa (H_1) $H1: \beta_i \neq 0$

Hipótesis en términos de Odds Ratio

Dado que:

$$OR_i = e^{\beta_i}$$

Las hipótesis equivalentes se formulan así:

- Hipótesis nula (H_0) $H0: OR_i = 1$
- Hipótesis alternativa (H_1) $H1: OR_i \neq 1$

Estas hipótesis se pondrían a prueba utilizando las estadísticas de Wald o el test de razón de verosimilitud.

Test de Wald

Evalúa si el estimador $\hat{\beta}_i$ está "suficientemente lejos" de 0 en relación con su error estándar.

$$Z = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

Bajo H_0, este estadístico sigue aproximadamente una distribución normal estándar.

Suele elevarse al cuadrado:

$$W = \left(\frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}\right)^2$$

que sigue una distribución chi-cuadrado con 1 grado de libertad.

- Si el p-valor < $\alpha \rightarrow$ se rechaza H_0
- Equivalente a verificar si el IC 95% del OR incluye o no el valor 1

Test de razón de verosimilitud

Compara dos modelos:

- Modelo reducido (sin la variable de interés)
- Modelo completo (incluye la variable)

Evalúa si agregar la variable mejora significativamente el ajuste.

$$LR = -2[logL(modelo\ reducido) - logL(modelo\ completo)]$$

Equivalente a la diferencia de deviances.

Bajo H_0, sigue aproximadamente una chi-cuadrado con grados de libertad igual al número de parámetros añadidos.

- Si el modelo completo mejora significativamente el ajuste $\rightarrow$ se rechaza H_0

Ejemplo práctico

En la primera parte de este módulo, ajustaremos un **modelo de asociación** entre uno o más factores de exposición (variables independientes) y un desenlace.

Para este ejemplo práctico, usaremos la base de datos **icu** descrita en el libro *"Applied Logistic Regression"* (Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. John Wiley & Sons). La base la encontrarán en la librería de R `aplore3` (como se explicará a continuación).

Esta base de datos corresponde a información de una muestra de 200 individuos de un estudio cuyo objetivo era evaluar la sobrevida de los pacientes ingresados en la Unidad de Cuidado Intensivo y estudiar los factores de riesgo asociados a la mortalidad. Los datos se recolectaron en el Baystate Medical Center, en Springfield (Massachusetts).

En la siguiente tabla, se presentan las variables medidas en este estudio:

Variable	Definición	Detalles / Códigos
id	Código de identificación	Número ID único del paciente
sta	Estado vital al egreso hospitalario	1: Vivo · 2: Fallecido
age	Edad	Años cumplidos
gender	Sexo	1: Masculino · 2: Femenino
race	Raza	1: Blanca · 2: Negra · 3: Otra
ser	Servicio al ingreso a UCI	1: Médico · 2: Quirúrgico
can	Cáncer como parte del problema actual	1: No · 2: Sí
crn	Antecedente de insuficiencia renal crónica	1: No · 2: Sí
inf	Infección probable al ingreso a UCI	1: No · 2: Sí
cpr	RCP previa al ingreso a UCI	1: No · 2: Sí
sys	Presión arterial sistólica al ingreso a UCI	mm Hg
hra	Frecuencia cardíaca al ingreso a UCI	Latidos por minuto

Variable	Definición	Detalles / Códigos
pre	Ingreso previo a UCI en los últimos 6 meses	1: No · 2: Sí
type	Tipo de ingreso	1: Electivo · 2: Urgencia
fra	Fractura (hueso largo, múltiple, cuello, área única o cadera)	1: No · 2: Sí
po2	PO_2 en gases arteriales iniciales	1: > 60 · 2: ≤ 60
ph	pH en gases arteriales iniciales	1: ≥ 7.25 · 2: < 7.25
pco	PCO_2 en gases arteriales iniciales	1: ≤ 45 · 2: > 45
bic	Bicarbonato en gases arteriales iniciales	1: ≥ 18 · 2: < 18
cre	Creatinina en gases arteriales iniciales	1: ≤ 2.0 · 2: > 2.0
loc	Nivel de conciencia al ingreso a UCI	1: Sin coma ni estupor profundo · 2: Estupor profundo · 3: Coma

Base de datos

Como primer paso, cargaremos los datos. La base está contenida en la librería `aplore3`. Para cargarla, únicamente debemos instalarla y llamarla a nuestro ambiente, asignándola al objeto `base_icu`, con el siguiente código:

```
install.packages("aplore3")
base_icu <- aplore3::icu
```

Además, haremos un corto resumen de nuestra base de datos para llevarnos una idea de las variables con las que estaremos trabajando:

```
?aplore3::icu
summary(base_icu)
##        id              sta             age              gender       race
##  Min.   :  4.0   Lived:160   Min.   :16.00   Male  :124   White:175
##  1st Qu.:210.2   Died : 40   1st Qu.:46.75   Female: 76   Black: 15
##  Median :412.5               Median :63.00                Other: 10
##  Mean   :444.8               Mean   :57.55
##  3rd Qu.:671.8               3rd Qu.:72.00
##  Max.   :929.0               Max.   :92.00
##        ser       can       crn       inf       cpr            sys
##  Medical : 93   No :180   No :181   No :116   No :187   Min.   : 36.0
##  Surgical:107   Yes: 20   Yes: 19   Yes: 84   Yes: 13   1st Qu.:110.0
##                                                         Median :130.0
##                                                         Mean   :132.3
##                                                         3rd Qu.:150.0
##                                                         Max.   :256.0
##       hra           pre              type      fra          po2             ph
##  Min.   : 39.00   No :170   Elective : 53   No :185   > 60 :184   >= 7.25:187
##  1st Qu.: 80.00   Yes: 30   Emergency:147   Yes: 15   <= 60: 16   < 7.25 : 13
##  Median : 96.00
##  Mean   : 98.92
##  3rd Qu.:118.25
##  Max.   :192.00
##     pco          bic           cre            loc
##  <= 45:180   >= 18:185   <= 2.0:190   Nothing:185
##  > 45 : 20   < 18 : 15   > 2.0 : 10   Stupor :  5
##                                       Coma   : 10
##
##
##
```

Encontramos que nuestra base de datos reune la información de 21 variables demográficas y clínicas registradas al momento del ingreso para cada uno de los 200 pacientes. El desenlace principal es el estado vital al alta hospitalaria (vivió o murió), con una **mortalidad del 20%**. La muestra tiene una edad promedio de 57.5 años, predominan los hombres y la mayoría de los ingresos fueron de emergencia. Además de antecedentes como cáncer, insuficiencia renal crónica o infección probable, la base incluye variables fisiológicas objetivas como presión arterial sistólica, frecuencia cardíaca, parámetros de gases arteriales y nivel de conciencia.

Formulación de la pregunta de investigación

Una vez comprendida la estructura general de la base y definido que nuestro desenlace es dicótomo (estado vital: vivió/murió), el siguiente paso consiste en formular claramente nuestra pregunta de investigación o el problema de modelamiento.

Nuestra pregunta de investigación será:

- **¿Existe una asociación significativa entre el servicio de procedencia del paciente que ingresa a la UCI y el estado vital del paciente al egreso?**

En este caso buscamos evaluar cuál es la relación entre una variable, considerada un factor de exposición (Variable `ser`: Servicio al ingreso a UCI), y una variable de desenlace cuya característica específica es que está medida en una escala cualitativa de tipo nominal, multinomial u ordinal (Variable `sta`: Estado vital al egreso hospitalario).

```
# Asignamos "Surgical" como nuestra categoría de referencia
base_icu$ser <- relevel(base_icu$ser, ref = "Surgical")
```

Adicionalmente, se tendrán en cuenta algunas variables que no son del interés primario, pero que pueden afectar la relación que se busca estudiar entre el factor de exposición y el desenlace. Estas variables constituyen las que llamaremos covariables o potenciales factores de confusión, y siguen teniendo el carácter de variables independientes.

Para el efecto de la especificación del modelo, tanto la variable de exposición como las covariables se denotarán como X_i y las denominaremos como variables independientes.

Supuestos del modelo (antes del ajuste)

Antes de ajustar el modelo, es fundamental realizar un análisis exploratorio que nos permita evaluar el cumplimiento de un par de condiciones - o supuestos - necesarias:

Supuesto 1: La variable dependiente debe ser binaria

- La **variable dependiente**, en este caso, `sta`, es de tipo cualitativo y dicótomo, con dos posibles valores: vivo (*Lived*) o muerto (*Died*).

```
install.packages("arsenal")
install.packages("dplyr")
base_icu %>%
  dplyr::select(sta) %>%
  arsenal::tableby( ~ ., data = ., digits = 1, test = TRUE) %>%
  summary()
```

	Overall (N=200)
sta	
Lived	160 (80.0%)
Died	40 (20.0%)

Cuando vamos a ajustar un modelo de regresión logística en R, debemos definir una variable desenlace (en este caso `sta`) y verificar sus categorías. Por defecto, R estimará la probabilidad de pertenecer a la segunda categoría que tengamos al aplicar la función `levels` (en este caso, morir (Died)).

```
levels(base_icu$sta)
```

[1] “Lived” “Died”

Supuesto 2: Supuesto de independencia

- Al revisar la base de datos, se observa que la información de cada individuo es independiente de la observación de cualquier otro. Por tanto, vemos que se cumple el supuesto de independencia en las observaciones.

Supuesto 3: Muestra suficiente

Con relación al supuesto de muestra suficiente y aunque este es un punto central en el caso de modelos de predicción, si se supone un **número de eventos por variable en el modelo de 10**, una frecuencia esperada del evento de 20% y 4 variables en el modelo, el tamaño de muestra de 200, resultaría suficiente para explorar la asociación de interés.

Supuesto 4: Multicolinealidad

Es importante evaluar si existen correlaciones altas entre variables independientes, ya que la colinealidad puede inflar los errores estándar, volver inestables los coeficientes y dificultar la interpretación del modelo. Una forma inicial y sencilla de explorar este problema es mediante la matriz de correlación, particularmente para variables continuas.

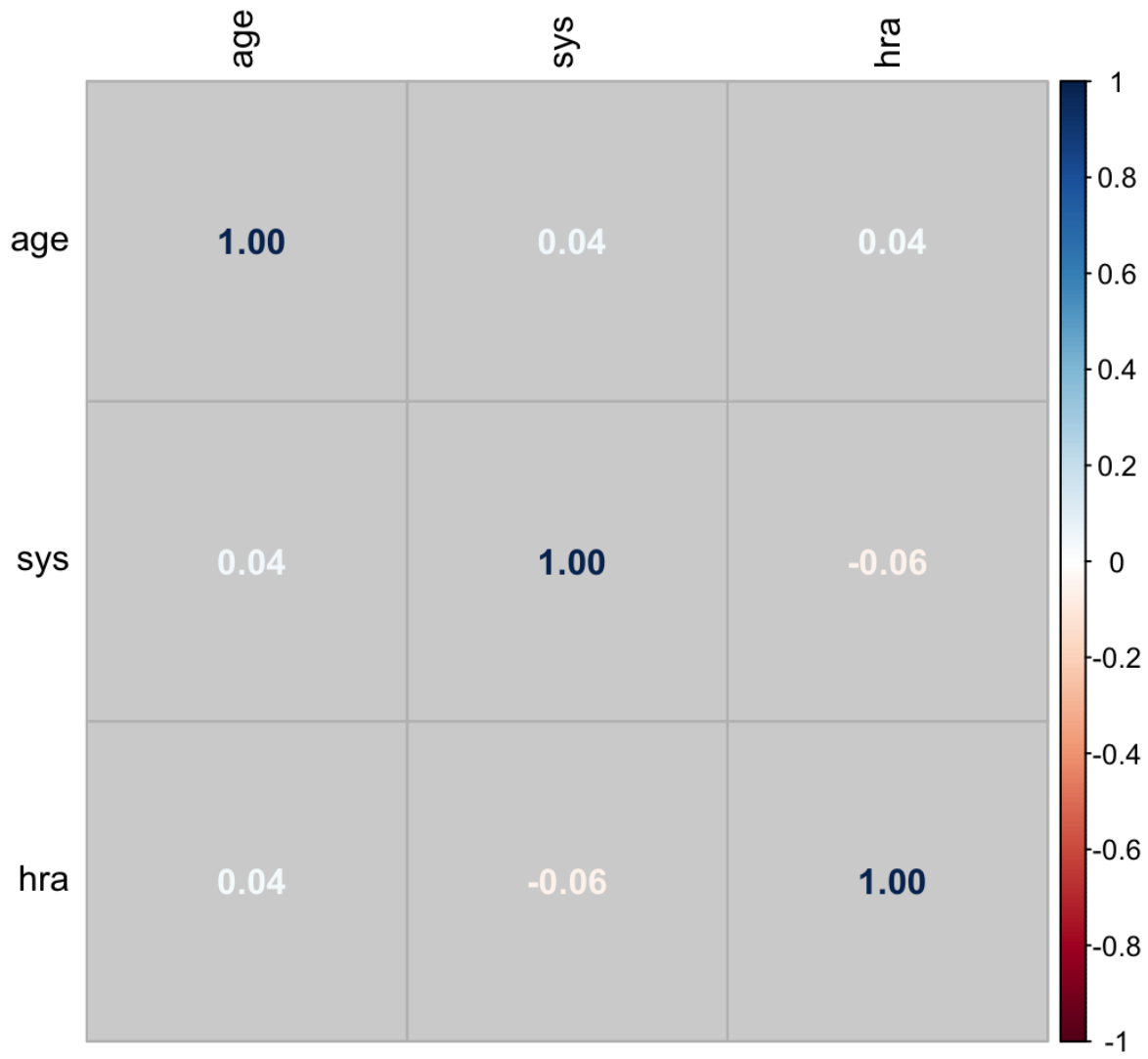
En nuestra base ICU, las variables continuas son:

- `age` (edad)
- `sys` (presión sistólica)
- `hra` (frecuencia cardíaca)

```
install.packages("corrplot")
# Seleccionamos las variables continuas
vars_cont <- base_icu[, c("age", "sys", "hra")]


# Calculamos la matriz de correlación de Pearson
cor_matrix <- cor(vars_cont, use = "complete.obs")


corrplot::corrplot(cor_matrix, method = "number", bg = "lightgrey", tl.col = "black")
```



La matriz de correlación muestra coeficientes entre -1 y 1:

- $r \approx 0 \rightarrow$ no hay correlación lineal
- $r > 0.7 \rightarrow$ posible colinealidad relevante
- $r > 0.8–0.9 \rightarrow$ colinealidad alta, potencialmente problemática

En nuestro ejemplo, todos los valores están por debajo de 0.7 por lo que podemos afirmar que no tenemos colinealidad entre nuestras variables independientes continuas.

Planteamiento del modelo

Una vez realizado el análisis exploratorio y definida la codificación del desenlace, el siguiente paso es especificar formalmente el modelo logístico. En regresión logística no modelamos directamente la probabilidad de morir, sino el logaritmo de las odds (logit) de ese evento. Es decir, planteamos un modelo de la forma:

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \cdots + \beta_k X_k$$

donde p es la probabilidad de morir y $X_1, X_2, \ldots, X_k$ son las variables independientes seleccionadas (por ejemplo: servicio al ingreso a UCI, edad, presión sistólica, presencia de infección, tipo de admisión, etc.).

En esta etapa debemos tomar algunas decisiones importantes:

1. ¿Qué variables incluir inicialmente?

Podemos comenzar con un modelo completo basado en criterio clínico (incluyendo la variable de exposición de interés y todos los posibles confusores desde el punto de vista clínico, siempre y cuando el tamaño de muestra lo permita) o con un modelo más parsimonioso.

2. ¿Cómo tratar las variables continuas?

Variables como edad, presión sistólica o frecuencia cardíaca deben evaluarse respecto al supuesto de linealidad en el logit. Si no cumplen este supuesto, podrían transformarse o categorizarse.

3. Codificación de variables categóricas

Las variables binarias pueden incluirse directamente (0/1), mientras que variables con más de dos categorías requieren variables indicadoras (dummies), definiendo una categoría de referencia.

Ajuste del modelo en R

Para efectos de este ejemplo, comenzaremos con un modelo univariado que incluya únicamente nuestra exposición:

- Servicio al ingreso a UCI (`ser`)

El modelo se ajusta con el siguiente código:

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

- `glm()`: Es la función en R para ajustar modelos lineales generalizados.
- `sta ~ ser`: especifica la fórmula del modelo:
  - `sta` → variable dependiente
  - Lo que está a la derecha del símbolo ~ serán las variables independientes
- `family = binomial`: indica que:
  - La distribución del error es binomial
  - Se utiliza la función de enlace logit (por defecto)

Esto convierte el modelo en una regresión logística.

Resultados e interpretación del modelo

Para ver los resultados del modelo ajustado, utilizaremos la función `summary`:

```
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
```

- A priori, con la columna **Pr(>|z|)** podemos indentificar qué variables muestran una asociación estadísticamente significativa con nuestro desenlace. En este caso, sería:
  - Servicio médico (vs. quirúrgico) (p = 0.0101)
  - Este valor p (0.0101) proviene del test de Wald aplicado al coeficiente `serMedical` (Servicio al ingreso a UCI: Médico)
    - $H0: \beta serMedical = 0$
    - $H1: \beta serMedical \neq 0$
  - Esto sería equivalente a: $H0: e^{\beta} serMedical = OR = 1$
- La columna ***Estimate*** nos da los coeficientes en escala logarítmica (log-odds), por lo que para interpretarlos clínicamente debemos exponenciarlos (lo haremos a continuación)

¿Cómo exponenciar los coeficientes del modelo en R?

Podemos exponenciar los coeficientes del modelo en R de forma manual o utilizando la función `tbl_regression` del paquete `gtsummary`:

```
install.packages("gtsummary")
## Forma manual de exponenciar los coeficientes del modelo
exp(coef(modelo))
## (Intercept)  serMedical
##   0.1505376   2.5778252
## 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	2.58	1.27, 5.43	**0.010**

Abbreviations: CI = Confidence Interval, OR = Odds Ratio

Interpretación del modelo

Los pacientes del servicio médico tienen 2.58 veces el odds (OR crudo) del evento comparados con los pacientes del servicio quirúrgico.

El intervalo de confianza nos dice: Con 95% de confianza, el verdadero OR poblacional está entre 1.27 y 5.43.

Puntos clave:

- El intervalo no incluye 1; por tanto, la asociación es estadísticamente significativa
- El efecto podría ser:
  - tan bajo como 1.27 (aumento moderado del odds)
  - tan alto como 5.43 (aumento grande del odds)

El intervalo es relativamente amplio → posible imprecisión (tal vez tamaño muestral moderado).

¿Podemos entonces hablar de "riesgo"?

El Riesgo Relativo (RR), también llamado *Relative Risk* o *Risk Ratio*, se define como:

$$RR = \frac{Riesgo\ en\ expuestos}{Riesgo\ en\ no\ expuestos}$$

En términos probabilísticos:

$$RR = \frac{P(Y = 1|X = 1)}{P(Y = 1|X = 0)}$$

Es decir:

El RR compara directamente probabilidades (riesgos).

En nuestro ejemplo, calculamos el RR a través de una tabla de 2x2:

```
base_icu %>%
  dplyr::select(sta, ser) %>%
  arsenal::tableby(sta ~ ., data = ., digits = 1, test = FALSE, cat.stats="countrowpct") %>%
  summary()
```

	Lived (N=160)	Died (N=40)	Total (N=200)
ser			
Surgical	93 (86.9%)	14 (13.1%)	107 (100.0%)
Medical	67 (72.0%)	26 (28.0%)	93 (100.0%)

```
tabla_2x2 <- addmargins(table(base_icu$ser, base_icu$sta))
# Riesgo en Medical
riesgo_med <- tabla_2x2["Medical","Died"] / sum(tabla_2x2["Medical",])

# Riesgo en Surgical
riesgo_surg <- tabla_2x2["Surgical","Died"] / sum(tabla_2x2["Surgical",])

# RR real
RR_real <- riesgo_med / riesgo_surg
round(RR_real, 2)
## [1] 2.14
```

Interpretación:

- Los pacientes admitidos del servicio médico tienen 2.14 veces el riesgo de morir comparados con los pacientes del servicio quirúrgico.

¿Y entonces el OR?

El OR compara odds, no probabilidades.

Recordemos:

$$odds = \frac{P}{1-P}$$

Entonces el OR es:

$$OR = \frac{odds\ en\ expuestos}{odds\ en\ no\ expuestos}$$

Es decir:

$$OR = \frac{\frac{P_1}{1-P_1}}{\frac{P_0}{1-P_0}}$$

Donde:

- $P1 = P(Y = 1|X = 1)$
- $P0 = P(Y = 1|X = 0)$

¿Por qué la regresión logística estima OR y no RR?

Porque el modelo logístico modela:

$$log\left(\frac{P}{1-P}\right)$$

Es decir, modela el log odds.

Cuando exponenciamos un coeficiente β:

$$e^{\beta}$$

obtenemos un odds ratio ajustado.

El modelo nunca modela directamente:

$$log(P)$$

por eso no produce RR directamente.

¿Cuándo el OR se acerca al RR?

Cuando la incidencia del desenlace (evento de interés) es baja (menos del 10% o 5%), entonces:

- La probabilidad de que ocurra el evento (a / (a + b)) es aproximadamente igual a (a / b) (la "odd" del evento).
- De manera similar, (c / (c + d)) es aproximadamente igual a (c / d).

Matemáticamente:

Si la probabilidad del evento es pequeña, entonces a + b ≈ b y c + d ≈ d. Por lo tanto:

$$RR = \frac{a/(a+b)}{\mathrm{c}/(\mathrm{c}+\mathrm{d})} \approx \frac{a/b}{c/d} = OR$$

$$RR = \frac{a/(a+b)}{\mathrm{c}/(\mathrm{c}+\mathrm{d})} \approx \frac{a/b}{c/d} = OR$$

Si el desenlace que estás estudiando es poco frecuente, puedes usar el OR como una buena estimación del RR. Sin embargo, si el evento es común (incidencia alta), el OR tenderá a sobrestimar el RR.

Volviendo a nuestro ejemplo…

Con nuestro ejemplo, tenemos

- RR = 2.14
- OR crudo = 2.58

En este caso, OR es aproximadamente 20% mayor que el RR. Esto ocurre porque nuestro desenlace no es raro (*tenemos 20% de muertes*). No podríamos interpretar el OR como un riesgo.

Recordemos:

$$odds = \frac{P}{1 - P}$$

Cuando P aumenta: - El denominador $(1 - P)$ disminuye - El odds crece más rápido que la probabilidad

Por eso el OR tiende a “inflarse” respecto al RR.

Lecturas complementarias:

- Kleinbaum, D. G., & Klein, M. (2010). Logistic Regression: A Self-Learning Text (3rd ed.). Springer
- Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. John Wiley & Sons
- Kleinbaum, D. G., Kupper, L. L., & Muller, K. E. (1988). Applied regression analysis and other multivariable methods (2nd ed.). ChPTER 21. PWS-Kent Publishing Company