LECCIÓN 7: REGRESIÓN LOGÍSTICA BINOMIAL.

SAAVEDRA, P.

1. El problema de la confusión

En el primer capítulo se discutió el problema de la asociación entre la *diabetes mellitus* y la *hipertensión arterial*. Usando los datos del estudio de Telde, observamos entonces una fuerte asociación entre ambos factores (Odds-ratio = 5.059; IC-95% = [3.419 ; 7.484]). Por tanto, tal asociación obviamente existe, pero ello no significa que se explique necesariamente por una relación de causalidad. Sugerimos entonces que la asociación podría explicarse por el efecto de confusión de la *edad* y de la *resistencia a la insulina*. En esta lección introduciremos los modelos de regresión logística a través de los cuales concluiremos que entre personas de la misma edad y con el mismo nivel de resistencia a la insulina, los hipertensos no tienen mayor probabilidad de ser diabéticos que los normotensos. Ello significa que la referida asociación cruda se explica por el efecto de confusión de estas dos variables.

2. Estudio de Telde

Ilustramos esta lección con el estudio de Telde, cuyo propósito inicial fue identificar los factores asociados con la diabetes mellitus de tipo 2. En los siguientes apartados se resume el estudio:

- ***Objetivo***: Identificar los factores que se asocian con la diabetes mellitus de tipo 2.
- ***Diseño***: Estudio transversal en el que se incluyeron 1030 personas con 30 ó más años del municipio de Telde.
- ***Mediciones***. Edad, sexo, medidas antropométricas (peso, talla y perímetro abdominal), nivel de estudios, estilos de vida (sedentarismo, tabaquismo y consumo de alcohol), presiones arteriales diastólica y sistólica, glucemia basal y postcarga, insulinemia, HbA1c%, determinaciones lípídicas (colesterol total, LDL,

HDL, triglicéridos, apo-A, apo-B y Lp(a)), creatinina, homocisteína, Fibrinógeno, PAI-1, FVW, PCR y polimorfismos genéticos

- ***Resultados básicos***: El número de sujetos con diabetes mellitus de tipo 2 fue de 128 (12,43%), siendo la tasa de prevalencia ajustada por la población der Telde del 12,62% mientras que la ajustada por la población SEGI fue del 13.81%. En la tabla 1 se resumen algunas de las variables del estudio según diagnóstico o no de diabetes mellitus de tipo 2.

TABLA 1. Características de los sujetos del estudio según presencia o no de diabetes mellitus de tipo 2

	Diabetes Mellitus tipo 2		
	Si (N=128)	No (N=902)	P-valor
Edad (años)	58.9±10.4	46.5±11.3	< 0.001
Sexo masculino	74 (57.8)	374 (41.5)	< 0.001
Índice de masa corporal (kg/m2)	31.0±5.4	27.8±4.8	< 0.001
Obesidad abdominal (ATP-III)	107 (83.6)	485 (53.8)	< 0.001
Hipertensión arterial (OMS)	83 (64.8)	241 (26.7)	< 0.001
Ln(HOMA-IR)	0.40±0.70	1.43±0.78	< 0.001
Colesterol-LDL (mg/dL)	135.6±36.6	133.7±31.4	0.543
Ln(Triglicéridos) (mg/dL)	4.99±0.49	4.63±0.49	< 0.001

Nótese que la tasa de prevalencia de hipertensión arterial es notablemente superior en el grupo de diabéticos ($p < 0.001$). Ello podría explicarse, al menos parcialmente, por el hecho de que la edad media estimada de los diabéticos (58.9 años) es superior a la de los no diabéticos (46.5 años). De esta forma, cabe esperar que la tasa de prevalencia de hipertensión en el subconjunto de diabéticos, por ser éstos de mayor edad, sea superior a la del subconjunto de no diabéticos. No obstante, otros factores podrían explicar también la fuerte asociación entre la diabetes mellitus y la hipertensión. A lo largo de las siguientes secciones probaremos que esta asociación es atribuible al efecto de confusión de la *edad* y la *resistencia a la insulina*.

3. Asociación *diabetes mellitus* - *hipertensión arterial*

En orden a estimar la asociación entre la diabetes mellitus de tipo 2 (DM) y la hipertensión arterial (HTA), consideramos la tabla de contingencia 2×2 mostrada en la tabla 2.

Tabla 2. Tabla de contingencia diabetes mellitus x hipertensión arterial

		Diabetes mellitus	
		Si (E)	No (C)
		$n_E = 128$	$n_C = 902$
Hipertensión arterial (HTA)	Si	83 (64.8)	241 (26.7)
	No	45 (35.2)	661 (73.3)

De la tabla 2 se tiene:

$$\pi_E = \Pr(HTA \mid DM) \quad \Longrightarrow \quad \hat{\pi}_E = 83/128 \quad ; \quad 1 - \hat{\pi}_E = 45/128$$

$$\pi_C = \Pr(HTA \mid DM^C) \quad \Longrightarrow \quad \hat{\pi}_C = 241/902 \quad ; \quad 1 - \hat{\pi}_C = 661/902$$

Entonces, el estimador natural de la *odds-ratio* (cruda) es:

$$\hat{\omega} = \frac{\hat{\pi}_E\,(1 - \hat{\pi}_C)}{\hat{\pi}_C\,(1 - \hat{\pi}_E)} = \frac{(83/128) \times (661/902)}{(241/902) \times (45/128)} = \frac{83 \times 661}{241 \times 45} = 5.059$$

En el siguiente cuadro se muestra una salida del paquete R conteniendo la estimación de la odds-ratio junto con el correspondiente intervalo de confianza al 95%.

```
Epidemiological 2x2 Table Analysis

Input Matrix:
            Diabetes
Hipertension  Sí  No
          Sí  83 241
          No  45 661

Pearson Chi-Squared Statistic (Includes Yates' Continuity Correction): 73.809
```

```
Associated p.value for H0: There is no association between exposure and outcome vs. HA
p.value using Fisher's Exact Test (1 DF) : 0

Estimate of Odds Ratio: 5.059
95% Confidence Limits for true Odds Ratio are: [3.419, 7.484]

Estimate of Relative Risk (Cohort, Col1): 4.019
95% Confidence Limits for true Relative Risk are: [2.866, 5.636]

Estimate of Risk Difference (p1 - p2) in Cohort Studies: 0.192
95% Confidence Limits for Risk Difference: [0.139, 0.246]

Estimate of Risk Difference (p1 - p2) in Case Control Studies: 0.381
95% Confidence Limits for Risk Difference: [0.315, 0.447]

Note: Above Confidence Intervals employ a continuity correction.
```

Como ya se ha indicado, la asociación estimada entre la hipertensión arterial (HTA) y la *DM* (Odds-ratio = 5.059; IC-95% = [3.419 ; 7.484]) no indica necesariamente que exista una relación de causalidad entre ambos factores pues tal asociación podría explicarse por factores de confusión. En tal sentido, investigaremos a la edad y la resistencia a la insulina como posibles *confounding* (confusores). Para elucidar este problema utilizaremos los modelos de regresión logística.

4. Modelo de regresión logística binomial

Consideremos una variable binaria Y que indica la ***presencia*** (1) ó ***ausencia*** (0) de un cierto carácter (por ejemplo*, diabetes mellitus de tipo 2* o *respuesta favorable de una enfermedad a un tratamiento*), la cual se desea explicar por un conjunto de variables numéricas $X_1, ..., X_p$. Para este fin puede utilizarse el modelo de ***regresión logística binomial,*** el cual tiene la forma:

$$\Pr\left(Y = 1 \mid X_1, ..., X_n\right) = \frac{\exp\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right)}{1 + \exp\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right)}$$

En el modelo, $\beta_0,\ \beta_1, ..., \beta_p$ son ***parámetros*** que corresponden a los efectos de las variables $X_1, ..., X_p$ sobre la variable Y (*variable de respuesta* o *outcome*) y que en general, se estimarán utilizando un conjunto de datos de la forma:

$$\left\{\left(x_{i,1}, ..., x_{i,p}\,; y_i\right)\ :\ i = 1, ..., n\right\}$$

Aquí, n representa el número de sujetos incluidos en el estudio, $x_{i,j}$ la observación de la variable X_j en el i-ésimo sujeto e y_i, el valor (1 ó 0) de la respuesta.

Obsérvese que la definición del modelo garantiza la propiedad:

$$0 < \Pr\left(Y = 1 \mid X_1, ..., X_p\right) < 1$$

Sin pérdida de generalidad, consideremos un modelo logístico con una sola covariable:

$$\Pr\left(Y = 1 \mid X\right) = \frac{\exp\left(\beta_0 + \beta_1 X\right)}{1 + \exp\left(\beta_0 + \beta_1 X\right)}$$

Estimación de los parámetros del modelo. Para la estimación del modelo se requiere de un conjunto de datos de la forma:

$$\left\{\left(x_i, y_i\right)\ :\ i = 1, ..., n\right\}$$

A partir de estos datos pueden obtenerse las estimaciones de los parámetros desconocidos β_i, las cuales expresamos por $\hat{\beta}_i$ y sus correspondientes errores estándar $\mathrm{sd}\left(\hat{\beta}_i\right)$. De aquí, los intervalos de confianza al nivel $1 - \alpha$ se obtienen como:

$$\left[\hat{\beta}_i - z_{1-\alpha/2} \cdot \mathrm{sd}\left(\hat{\beta}_i\right)\ ;\ \hat{\beta}_i + z_{1-\alpha/2} \cdot \mathrm{sd}\left(\hat{\beta}_i\right)\right]$$

siendo $z_{1-\alpha/2}$ el cuantil $1 - \alpha/2$ de la distribución de probabilidad normal estándar $\left(N\left(0, 1\right)\right)$. Nótese que para $\alpha = 0.05$, $z_{1-\alpha/2} = z_{0.975} = 1.96$.

Contraste de asociación. Si $\beta_1 = 0$, las variaciones en el valor de X no influirían sobre la probabilidad $\Pr\left(Y = 1 \mid X\right)$, lo que significaría la inexistencia de asociación entre X e Y. Por tal motivo, el contraste de asociación tiene la forma: $H_0 : \beta_1 = 0$ frente a la hipótesis alternativa $H_1 : \beta_1 \neq 0$. Los paquetes estadísticos proporcionan entonces el p-valor para este contraste.

5. Asociación diabetes - hipertensión ajustada por edad

Tal como se ha señalado, la asociación identificada entre la DM y la HTA se podría explicar por el efecto de confusión de otras variables, tales como la edad. Para elucidar este problema, estimaremos la odds-ratio que mide la asociación entre DM y HTA entre ***sujetos de una misma edad***. Para ello usaremos el siguiente modelo de regresión logistica:

$$\Pr\left(DM = 1 \mid HTA, Edad\right) = \frac{\exp\left(\beta_0 + \beta_1 Edad + \beta_2 HTA\right)}{1 + \exp\left(\beta_0 + \beta_1 Edad + \beta_2 HTA\right)}$$

Aquí, las variables DM yHTA se codifican como 1 ó 0 según el correspondiente carácter esté presente o no. En la tabla 3 se muestra la estimación del modelo utilizando los datos del estudio de Telde.

Tabla 3. Estimación del modelo logístico para la diabetes mellitus de tipo 2 (DM) siendo las covariables la edad y la hipertensión arterial (HTA) basada en los datos del estudio de Telde)

	Estimación	Error estándar	P-valor
β_0 (término independiente)	-6.195	0.511	< 0.001
β_1 (*Edad)*	0.073	0.009	< 0.001
β_2 (*Hipertensión arterial*)	0.944	0.220	< 0.001

De acuerdo con este modelo, a una misma edad puede observarse que la probabilidad de hipertensión arterial difiere entre diabéticos y no diabéticos ($p < 0.001$). Este es un hecho que puede apreciarse en la figura 1

Odds-ratio ajustada por edad. Las probabilidades que se muestran en la tabla 4 se han obtenido del modelo logístico para una misma edad **fija**.

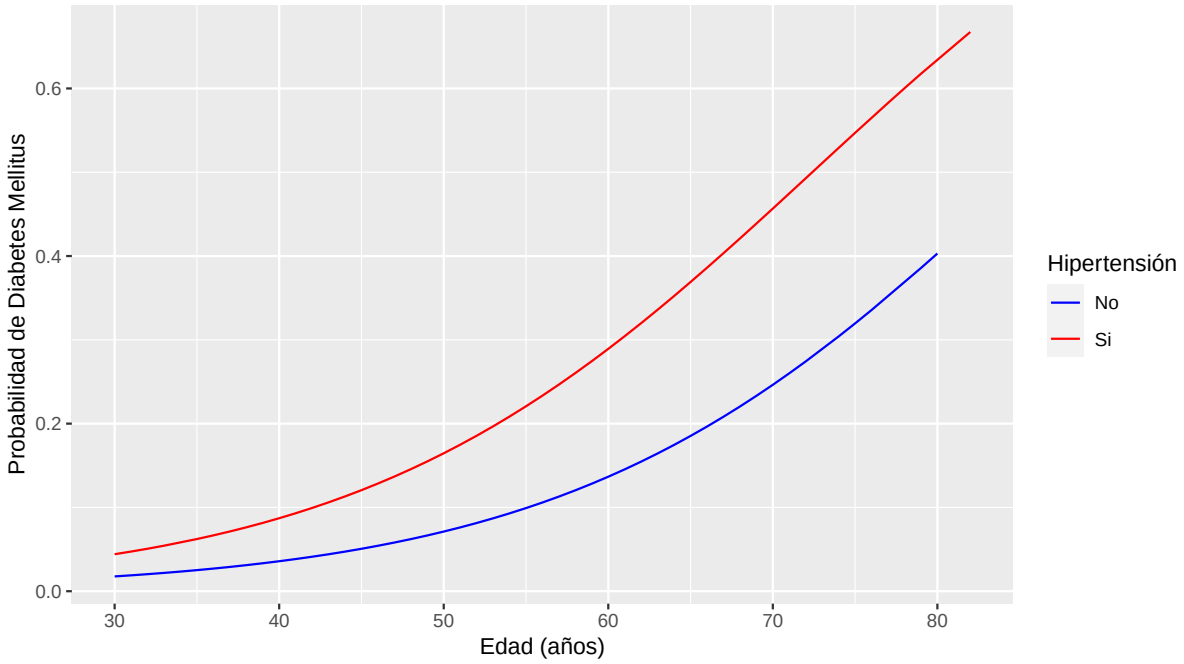

FIGURA 1. Evolución de las probabilidades estimadas de diabetes mellitus según presencia o no de hipertensión arterial

TABLA 4. Tabla de contingencia para diabetes (DM) e hipertensión arterial (HTA) en pacientes con una misma edad. La presencia/ausencia de cada carácter se codifica como 1/0

	$DM = 1$	$DM = 0$
$HTA = 1$	$\Pr(DM = 1 \mid Edad, HTA = 1)$	$\Pr(DM = 0 \mid Edad, HTA = 1)$
$HTA = 0$	$\Pr(DM = 1 \mid Edad, HTA = 0)$	$\Pr(DM = 0 \mid Edad, HTA = 0)$

De la tabla anterior, podemos obtener la odds-ratio que mide la asociación entre DM e HTA entre sujetos de una misma edad en la forma:

$$\frac{\Pr(DM = 1 \mid Edad, HTA = 1) \times \Pr(DM = 0 \mid Edad, HTA = 0)}{\Pr(DM = 1 \mid Edad, HTA = 0) \times \Pr(DM = 0 \mid Edad, HTA = 1)} = \exp(\beta_2)$$

Simples operaciones aritméticas permiten comprobar que el cociente anterior es $\exp(\beta_2)$. Esta cantidad es por definición la ***odds-ratio*** que mide la asociación entre DM e HTA ***ajustada por edad*** (entre sujetos de una misma edad).

Dado que $\left[\hat{\beta}_2 - z_{1-\alpha/2} \cdot \text{sd}\left(\hat{\beta}_2\right) \; ; \; \hat{\beta}_2 + z_{1-\alpha/2} \cdot \text{sd}\left(\hat{\beta}_2\right)\right]$ es un intervalo de confianza al nivel $1-\alpha$ para β_2, el intervalo de confianza a ese nivel para la odds-ratio ajustada es:

$$\left[\exp\left(\hat{\beta}_2 - z_{1-\alpha/2} \cdot \text{sd}\left(\hat{\beta}_2\right)\right) \; ; \; \exp\left(\hat{\beta}_2 + z_{1-\alpha/2} \cdot \text{sd}\left(\hat{\beta}_2\right)\right)\right]$$

La tabla 5 muestra los intervalos de confianza obtenidos para las odds-ratios ajustadas en el modelo anterior.

TABLA 5. Estimación del modelo logístico para la diabetes mellitus de tipo 2 (DM) siendo las covariables la edad y la hipertensión arterial (HTA) basada en los datos del estudio de Telde)

	Coeficiente (SE)	P-valor	Odds-ratio (IC-95%)
β_0	-6.195 (0.511)	< 0.001	-
Edad (años)	0.073 (0.009)	< 0.001	1.075 [1.056 ; 1.095]
Hipertensión arterial	0.944 (0.220)	< 0.001	2.571 [1.670 ; 3.958]

EJERCICIO

La resistencia a la insulina se evalúa a través del marcador HOMA-IR (por sus siglas en inglés *Homeostatic Model Assessment for Insulin Resistance*) el cual es proporcional al producto de la *glucemia basal* por la *insulinemia basal*. Nótese que la concurrencia de una elevada glucemia basal con una elevada insulinemia basal es una clara consecuencia de la resistencia a la insulina. En la tabla 6 se muestra la estimación del modelo logístico:

$$\Pr\left(DM = 1 \mid HTA, Edad, LnHOMA\right) = \frac{\exp\left(\beta_0 + \beta_1 Edad + \beta_2 HTA + \beta_3 LnHOMA\right)}{1 + \exp\left(\beta_0 + \beta_1 Edad + \beta_2 HTA + \beta_3 LnHOMA\right)}$$

Aquí, $LnHOMA = \ln\left(HOMA\right)$.

TABLA 6. Estimación del modelo logístico para la diabetes mellitus de tipo 2 (*DM*) siendo las covariables la edad, hipertensión arterial (*HTA*) y el HOMA (en escala logarítmica) basada en los datos del estudio de Telde)

	Coeficiente (SE)	*P*-valor	Odds-ratio (IC-95%)
β_0	-8.132 (0.680	< 0.001	-
Edad (por año)	0.079 (0.011)	< 0.001	1.082 [1.059 ; 1.105]
Hipertensión arterial	0.291 (0.249)	0.243	1.338 [0.821 ; 2.181]
Ln(*HOMA-IR*) (por unidad)	2.064 (0.211)	< 0.001	7.874 [5.211 ; 11.90]

¿Puede afirmarse que exista una relación de causalidad entre la diabetes mellitus de tipo 2 y la hipertensión arterial? Hacer una discusión usando los resultados mostrados anteriormente en esta lección.

REFERENCIAS

1. Boronat M, Varillas VF, Saavedra P, Suárez V, Bosch E, Carrillo A, Nóvoa FJ. Diabetes mellitus and impaired glucose regulation in the Canary Islands (Spain): prevalence and associated factors in the adult population of Telde, Gran Canaria. Diabet Med. 2006 Feb;23(2):148-55. doi: 10.1111/j.1464-5491.2005.01739.x. PMID: 16433712.