Uso de métodos logísticos para identificar asociaciones entre Esquizofrenia y Fumar

Hernando Manuel Quintana Ávila

Universidad EAFIT
Sede Medellín
Facultad de Ciencias y Humanidades
Posgrado en Matemáticas Aplicadas
2013

Uso de métodos logísticos para identificar asociaciones entre Esquizofrenia y Fumar

Hernando Manuel Quintana Ávila

Tesis de grado presentada para optar al título de Magister en Matemáticas Aplicadas

Directores:
Wbaldo Londoño MS.
Francisco Javier Díaz Ceballos, Ph.D (Externo)

Universidad EAFIT
Sede Medellín
Facultad de Ciencias y Humanidades
Posgrado en Matemáticas Aplicadas
2013

Índice general

Dedicatoria

A mi esposa Diana María y mi hijo Christian por su apoyo permanente.

Agradecimientos

Especialmente por su acompañamiento e información suministrada a:
José de León, M.D.
Mental Health Research Center at Eastern State Hospital, Lexington, KY, USA.
Manuel Gurpegui, M.D.
M. Carmen Aguilar, M.D.
José M. Martínez-Ortega, M.D.
Department of Psychiatry and Institute of Neuroscience, University of Granada, Granada, España.
Mi más sincero reconocimiento a los docentes del Post-grado que contribuyeron en mi mejoramiento académico y profesional.
Un reconocimiento especial a Jonathan Taborda, por la diagramación y diseño del Texto con LaTeX.

Capítulo 1

Introducción

La literatura esta llena de artículos sobre la asociación entre fumar y esquizofrenia. Una búsqueda sobre «Esquizofrenia y (fumar o nicotina)» en Julio de 2011 produjo 1341 artículos. Si no se cuenta con la asesoría de un experto en esta área, se pensaría que más de los 1000 artículos publicados dan una buena comprensión o entendimiento sobre esta asociación. Desafortunadamente, este no es el caso. Muchos de los artículos publicados incluyen afirmaciones simplistas debido a la mala interpretación de asociaciones estadísticas y relaciones causales.

Los Clínicos y de ciencias básicas pueden no estar al tanto de las limitaciones del método epidemiológico y no darse cuenta que muchas de las afirmaciones publicadas en artículos sobre la asociación entre esquizofrenia y fumar son erróneas, según principios epidemiológicos básicos. De hecho es posible que muchos de los artículos sobre esta asociación publicados en los 90′s en revistas psiquiátricas no serían hoy aceptadas para publicación en revistas epidemiológicas como fueron originalmente publicadas.

Muchos de estos artículos fallaron al no tener en cuenta la imposibilidad de establecer relaciones de causalidad de estudios cruzados y no corrigieron las asociaciones estadísticas reportadas para confundir efectos usando estadística multivariada.

Otro problema con la información proporcionada en la sección de revisión de artículos publicados sobre esquizofrenia y fumar es que algunos autores frecuentemente tienden a ignorar estudios que no cuadran con sus hipótesis y no dan la debida importancia a replicar consistentemente resultados cuando se describe la literatura como se afirma en Ioannidis (2005). La literatura medica está llena de hallazgos falsos debido a la publicación de efectos menores pero significativos observados en muestras pequeñas.

El factor más importante en establecer un avance científico en medicina es la replicación del hallazgo a través de muchas muestras independientes. En particular, es necesario que el hallazgo sobreviva las variaciones típicamente asociadas con el «ruido» presente en investigación clínica. A este respecto estamos convencidos de que la asociación entre esquizofrenia y fumar es un hallazgo «verdadero» ya que ha sido replicado consistentemente en todo el mundo (de León, Díaz et al. 2005).

Sin embargo no se puede asegurar que muchos resultados de artículos publicados en el contexto de esquizofrenia y fumar son confiables por lo que solo unos pocos han sido replicados. Así, afirmaciones simplistas como «fumar aumentan los síntomas negativos o efectos colaterales antipsicóticos en pacientes de esquizofrenia» no son apoyados por una revisión exhaustiva de la literatura limitada disponible (de León y otros 2006).

En este trabajo se usan dos medidas estadísticas para explorar asociaciones: odds ratios ajustados (ORs) y riesgos atribuibles a población (PAR) calculados con estos *ORs* (Bolton y Robinson 2010; Greenland and Drescher 1993). Los *OR* es una medida del poder de asociación entre dos variables dicótomas: A más diferente de 1 un *ORs* es mas fuerte esa asociación. Cuando los intervalos de confianza a 95 % (CIs) de dos *ORS* no se interceptan se consideran significativamente diferentes. Un método multivariado, regresión logística, permite ajustar *ORs*, para variables potencialmente confusas. Una vez que se ajusta OR, si se asume que la *ORs* mide una relación causal entre una condición y un factor de riesgo no explicado por factores de confusión, los estimados PAR estiman la reducción porcentual en la prevalencia de la condición que seria observada si el factor de riesgo fuera removido de la población.

En el capítulo 2 se presenta el marco teórico de las variables binarias y regresiones logísticas como un caso de modelos lineales generalizados en los cuales las variables resultado, respuesta, se miden en escala binaria. En el capítulo 3 del presente trabajo se encuentra una descripción de los conceptos claves utilizados para explorar asociaciones: riesgos y riesgos relativos, odds y cocientes de odds. En el capítulo 4 se describe como utilizar e interpretar los coeficientes estimados de regresión logística en el cálculo de cocientes de odds (Odds Ratio) o simplemente *ORs*, usados para explorar asociaciones; como también los riesgos atribuibles a población (PAR) calculados con estos *ORs*. En el

capìtulo 5 se hace una descripción del diseño de estudio para explorar el comportamiento de fumar y esquizofrenia . En el capítulo 6 se presentan los resultados de dos aplicaciones publicadas.

1. *Smoking initiation and schizophrenia: a replication study in a Spanish sample*.

2. *Fewer but heavier caffeine consumers in schizophrenia: A case-control study*.

Capítulo 2

Variables Binarias y Regresión Logística

2.1. Distribuciones de Probabilidad

En este capítulo se consideran modelos lineales generalizados en los cuales las variables resultado, respuesta, que será definida muchas veces como Y se miden en escala binaria, parametrizadas en términos de $1|0$. Por ejemplo las respuestas pueden ser «muerto o vivo», o «presente o ausente». «Éxito» y «Falla» se usan como términos genéricos para las dos categorías. Uno (1) tradicionalmente indica un éxito; cero (0) como falla o no éxito. Uno también puede mirarse como poseyendo alguna propiedad, condición o característica y cero como carecer de la propiedad, condición o característica.
Una variable aleatoria Z es binaria si:

$$Z = \begin{cases} 1 & \text{si el resultado es éxito} \\ 0 & \text{si el resultado es falla} \end{cases}$$

Con

$$Pr\left(Z = 1\right) = \pi \quad \text{y} \quad Pr\left(Z = 0\right) = 1 - \pi.$$

Si hay n variables aleatorias $Z_1, Z_2, \cdots, Z_n$ que son independientes con $Pr\left(Z_j = 1\right) = \pi_j$, entonces la probabilidad conjunta es:

$$\prod_{j=1}^{n} \pi_j^{Z_j} \left(1 - \pi_j\right)^{1 - Z_j} = \exp\left[\sum_{j=1}^{n} Z_j \log\left(\frac{\pi_j}{1 - \pi_j}\right) + \sum_{j=1}^{n} \log\left(1 - \pi_j\right)\right] \tag{2.1}$$

La cual es un miembro de la familia exponencial.

2.2. Modelos lineales generalizados

La unidad de muchos métodos estadísticos que implican combinaciones lineales de los parámetros fue demostrada por Nelder y Wedderburn (1972) usando la idea de modelos lineales generalizados. Esto se define en términos de un conjunto de variables aleatorias independientes $Y_1, \cdots, Y_N$ cada una con una distribución de la familia exponencial con las siguientes propiedades:

1. La distribución de cada Y_i es de la forma canónica y depende de un solo parámetro θ_i (los parámetros θ_i' s no tienen por qué ser el mismo) así

$$f\left(y_i; \theta_i\right) = \exp\left[y_i b_i(\theta_i) + c_i(\theta_i) + d_i(y_i)\right].$$

2. Las distribuciones de todas las Y_i's son de la misma forma (por ejemplo todas normales o todas binomiales) de modo que los subíndices en b, c y d no son necesarios.

Así, la función de densidad de probabilidad conjunta de $Y_1, \cdots, Y_N$ es

$$f(y_1, \cdots, y_N; \theta_1, \cdots, \theta_N) = \exp\left[\sum_{i=1}^{N} y_i b(\theta_i) + \sum_{i=1}^{N} c(\theta_i) + \sum_{i=1}^{N} d(y_i)\right] \tag{2.2}$$

Para la especificación del modelo, los parámetros θ_i no son por lo general de interés directo (ya que puede haber uno para cada observación). Para un modelo lineal generalizado se considera un conjunto más pequeño de los parámetros $\beta_1, \cdots, \beta_p$ (donde $p < N$) de tal manera que una combinación lineal de los parámetros β's es igual a alguna función del valor esperado μ_i de Y_i , es decir

$$g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

Donde g es una función monótona diferenciable llamada **función de enlace**.
$\boldsymbol{x_i}$: es un vector $p \times 1$, de variables explicativas (covariables y variables ficticias para los niveles de los factores).
$\boldsymbol{\beta}$: es un vector $p \times 1$, de parámetros

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

Por lo tanto un modelo lineal generalizado tiene tres componentes:

i) Las variables $Y_1, \cdots, Y_N$ de respuesta, que se supone que comparten la misma distribución de la familia exponencial.

ii) Un conjunto de parámetros $\boldsymbol{\beta}$ y las variables explicativas

$$\mathbf{X} = \begin{bmatrix} \boldsymbol{x_1^T} \\ \vdots \\ \boldsymbol{x_N^T} \end{bmatrix}$$

iii) Una función monótona de enlace g tal que

$$g(\mu_i) = \boldsymbol{x_i^T \beta}$$

Donde

$$\mu_i = E(Y_i)$$

Un caso especial es el modelo lineal

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$$

Donde los elementos e_i de $\boldsymbol{e}$ son independientes y todos tienen la distribución $N\left(0, \sigma^2\right)$. Este es un modelo lineal generalizado porque los elementos de $\boldsymbol{y}$ son variables aleatorias independientes Y_i con distribución $N\left(\mu_i, \sigma^2\right)$ donde $\mu_i = \boldsymbol{x_i^T \beta}$, la distribución normal es un miembro de la familia exponencial (siempre σ^2 se considera conocido) y, en este caso, g es la función identidad, esto es, $g(\mu_i) = \mu_i$.
Ahora si para n variables aleatorias $Z_1, Z_2, \cdots, Z_n$ binarias consideradas en la sección [2.1], que son independientes, las $\pi_j's$ son totas iguales, se puede definir

$$Y = \sum_{j=1}^{n} Z_j$$

O sea tal que Y es el número de éxitos en n ensayos. La variable aleatoria Y tiene la distribución binomial $b(n, \pi)$

$$Pr\left(Y = y\right) = \binom{n}{y} \pi^y \left(1 - \pi\right)^{n-y} \qquad y = 0, 1, 2, \cdots, n \tag{2.3}$$

Finalmente se considera el caso general de N variables aleatorias independientes $Y_1, Y_2, \cdots, Y_N$ correspondientes a los números ó cantidades de éxito en N subgrupos diferentes ó estratos [Cuadro 2.1]. Si $Y_i \sim b(n_i, \pi_i)$ la función de log-verosimilitud es

$$l(\pi_1, \pi_2, \cdots, \pi_N; y_1, y_2, \cdots, y_N) = \left[\sum_{i=1}^{N} y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \log(1-\pi_i) + \log\binom{n_i}{y_i} \right] \tag{2.4}$$

La distribución [2.4] no corresponde directamente a la ecuación [2.2] para la familia exponencial porque los $n_i's$ pueden no ser los mismos. Sin embargo, si la distribución conjunta de los $Y_i's$ se escribe en términos de las variables binarias Z_i, se sigue de [2.1] que [2.4] si pertenece a la familia de distribuciones exponenciales.

	Subgrupos			
	1	2	.....	N
Éxitos	Y_1	Y_2	.....	Y_N
Fracasos	$n_1 - Y_1$	$n_2 - Y_2$	.....	$n_N - Y_N$
Totales	n_1	n_2	.....	n_N

Cuadro 2.1: Frecuencias para N distribuciones binomiales

Se desea describir la proporción de éxitos $P_i = Y_i/n_i$, en cada subgrupo en términos de niveles de factor y otras variables explicativas las cuales caracterizan el subgrupo. Esto se hace modelando las probabilidades π_i como

$$g(\mu_i) = \boldsymbol{x_i^T \beta}$$

Donde $\boldsymbol{x_i}$ es un vector de variables explicativas (variables mudas para niveles de factor y valores medidos de covariados), $\boldsymbol{\beta}$ es un vector de parámetros y g es una función monótona, diferenciable llamada **función de enlace**.
El caso más simple es el **modelo lineal**

$$\pi = \boldsymbol{x^T \beta}$$

que se usa en algunas aplicaciones prácticas, pero tiene la desventaja de que aunque π es una probabilidad los valores ajustados $\boldsymbol{x^T b}$ pueden caer fuera del intervalo $[0, 1]$.
Para asegurar que π se restringe al intervalo $[0, 1]$, a menudo se modela usando una distribución de probabilidad acumulada

$$\pi = g^{-1}(\boldsymbol{x^T \beta}) = \int_{-\infty}^{t} f(s)ds$$

donde $f(s) \geq 0$ y $\int_{-\infty}^{\infty} f(s)ds = 1$. La función de densidad de probabilidad $f(s)$ se llama **distribución de tolerancia.**

2.3. Métodos de estimación de parámetros

Dos de los métodos más comunes utilizados para la estimación estadística de los parámetros son el **método de máxima verosimilitud** y el **método de mínimos cuadrados**. En esta sección se revisa el principio de cada uno de estos métodos y algunas propiedades de los estimadores. A continuación, el método de máxima verosimilitud se utiliza para la estimación de los parámetros de modelos lineales generalizados.
Por lo general las estimaciones deben ser obtenidas numéricamente por un procedimiento iterativo que resulta estar estrechamente relacionados con estimación ponderada de mínimos cuadrados.

2.3.1. Método de máxima verosimilitud

Sean $Y_1, Y_2, \cdots, Y_N$, N variables con función de densidad de probabilidad conjunta

$$f(y_1, \cdots, y_N; \theta_1, \cdots, \theta_p)$$

las cuales dependen de los parámetros $\theta_1, \cdots, \theta_p$. Por brevedad se denota

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \quad \text{y} \quad \boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix}$$

Por lo tanto la función de densidad conjunta es denotada por $f(\boldsymbol{y}; \boldsymbol{\theta})$.
La **función de verosimilitud** $L(\boldsymbol{\theta}; \boldsymbol{y})$ es algebraicamente igual que $f(\boldsymbol{y}; \boldsymbol{\theta})$ pero el cambio en notación refleja un cambio de énfasis de la variable aleatoria $\boldsymbol{y}$, con $\boldsymbol{\theta}$ fijo, a los parámetros $\boldsymbol{\theta}$ con $\boldsymbol{y}$ fijo (donde $\boldsymbol{y}$ representa las observaciones).
Sea Ω el conjunto de todos los posibles valores del vector de parámetros $\boldsymbol{\theta}$ (Ω se llama el espacio de parámetros). El **estimador de máxima verosimilitud** de $\boldsymbol{\theta}$ es el valor $\widehat{\boldsymbol{\theta}}$ el cual maximiza la función de verosimilitud, esto es

$$L\left(\widehat{\boldsymbol{\theta}}; \boldsymbol{y}\right) \geq L\left(\boldsymbol{\theta}; \boldsymbol{y}\right) \qquad \forall \boldsymbol{\theta} \in \Omega$$

De manera equivalente, $\widehat{\boldsymbol{\theta}}$ es el valor que maximiza la función log-verosimilitud $l(\boldsymbol{\theta}; \boldsymbol{y}) = \log L(\boldsymbol{\theta}; \boldsymbol{y})$ (puesto que la función logarítmica es monótona). Así

$$l(\widehat{\boldsymbol{\theta}}; \boldsymbol{y}) \geq l(\boldsymbol{\theta}; \boldsymbol{y}) \qquad \forall \boldsymbol{\theta} \in \Omega$$

A menudo es más fácil trabajar con la función log-verosimilitud que con la función verosimilitud.
En general el estimador $\widehat{\boldsymbol{\theta}}$ se obtiene diferenciando la función log-verosimilitud con respecto a cada elemento θ_i de $\boldsymbol{\theta}$ y resolviendo simultáneamente el sistema de ecuaciones

$$\frac{\partial l(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \theta_i} = 0 \quad \text{para} \quad i = 1, 2, \cdots, p$$

Para comprobar que las soluciones corresponden a un máximo de $l(\boldsymbol{\theta}; \boldsymbol{y})$ se verifica que la matriz de segundas derivadas

$$\frac{\partial^2 l(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \theta_j \theta_k}$$

Evaluada en $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ es definida negativa (esto es, si solo hay un parámetro θ para probar que

$$\frac{\partial^2 l(\theta; \boldsymbol{y})}{\partial \theta^2}$$

evaluada en $\theta = \widehat{\theta}$ es negativa).
También es necesario comprobar si hay algún valor de $\boldsymbol{\theta}$ en los bordes del espacio de parámetros Ω que son máximos locales de $l(\boldsymbol{\theta}; \boldsymbol{y})$. El estimador de máxima verosimilitud $\widehat{\boldsymbol{\theta}}$ corresponde al mayor de los máximos locales identificados.

Una propiedad importante de estimadores de máxima verosimilitud (a veces llamada propiedad de invariancia) es que si $g(\boldsymbol{\theta})$ es una función de los parámetros $\boldsymbol{\theta}$, entonces el estimador de máxima verosimilitud de $g(\boldsymbol{\theta})$ es $g(\widehat{\boldsymbol{\theta}})$. Una consecuencia es que podemos usar cualquier función conveniente de los parámetros para la estimación de máxima verosimilitud y luego usar la propiedad de invariancia para obtener estimaciones de máxima verosimilitud de los parámetros requeridos.
Otras propiedades de los estimadores de máxima verosimilitud son la consistencia, suficiencia y eficiencia asintótica.

2.3.2. Método de mínimos cuadrados

Sean $Y_1, Y_2, \cdots, Y_N$ variables aleatorias con valores esperados

$$E(Y_i) = \mu_i \quad \text{para} \quad i = 1, \cdots, N$$

Sean las $\mu_i's$ funciones de los parámetros $\beta_1, \cdots, \beta_p$ (donde $p < N$) que serán estimados. Sea

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

Consideremos la formulación

$$Y_i = \mu_i + e_i \quad \text{para} \quad i = 1, \cdots, N$$

En la cual μ_i representa la componente de «señal» para Y_i y e_i representa la componente «ruido».

El método de mínimos cuadrados consiste en encontrar estimadores $\widehat{\boldsymbol{\beta}}$, también denotados por $\boldsymbol{b}$, que minimice la suma de los cuadrados de los errores e_i; es decir minimizar la función

$$S = \sum e_i^2 = \sum [Y_i - \mu_i(\boldsymbol{\beta})]^2 \tag{2.5}$$

En notación matricial sería

$$S = (\boldsymbol{y} - \boldsymbol{\mu})^T (\boldsymbol{y} - \boldsymbol{\mu})$$

Donde

$$\boldsymbol{y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix} \quad \text{y} \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_N \end{bmatrix}$$

Generalmente el estimador $\widehat{\boldsymbol{\beta}}$ es obtenido diferenciando S con respecto a cada elemento β_j de $\boldsymbol{\beta}$ y resolviendo simultáneamente las ecuaciones

$$\frac{\partial S}{\partial \beta_j} = 0 \quad j = 1, \cdots, p$$

Por supuesto es necesario probar que las soluciones corresponden a los mínimos (es decir la matriz de segundas derivadas es definida positiva) e identificar el mínimo global de entre estas soluciones y cualquier mínimo local en los límites del espacio de parámetros.
En la práctica puede haber información adicional de las $Y_i's$; por ejemplo que algunas observaciones son menos fiables (es decir tienen una mayor varianza) que otras. En tales casos se puede ponderar los términos de la suma, quedando la suma para minimizar de la forma

$$S_W = \sum W_i [Y_i - \mu_i(\boldsymbol{\beta})]^2$$

Donde los términos W_i representan los pesos, por ejemplo, $W_i = [VarY_i]^{-1}$
De manera más general las $Y_i's$ pueden estar correlacionadas; sea $\boldsymbol{V}$ su matriz de varianza-covarianza. Luego los estimadores de **mínimos cuadrados ponderados** se obtienen minimizando

$$S_W = (\boldsymbol{y} - \boldsymbol{\mu})^T \boldsymbol{V^{-1}} (\boldsymbol{y} - \boldsymbol{\mu})$$

En particular si los términos μ_i son combinación lineal de los parámetros $\beta_j (j = 1, \cdots, p \quad \text{donde} p < N)$ esto es, si $\boldsymbol{\mu} = \boldsymbol{X\beta}$ es una matriz de orden $N \times p$, entonces

$$S_W = (\boldsymbol{y} - \boldsymbol{X\beta})^T \boldsymbol{V^{-1}} (\boldsymbol{y} - \boldsymbol{X\beta}) \tag{2.6}$$

Las derivadas de S_W con respecto a los elementos β_j de $\boldsymbol{\beta}$ son el vector

$$\frac{\partial S_W}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X\beta})$$

Por lo que los mínimos cuadrados ponderados $\boldsymbol{b}$ del vector de parámetro $\boldsymbol{\beta}$ es la solución de las **Ecuaciones normales**

$$\boldsymbol{X^T V^{-1} X b} = \boldsymbol{X^T V^{-1} y} \tag{2.7}$$

(ya que también se puede demostrar que la matriz de segundas derivadas es definida positiva.)

Una diferencia importante entre los métodos de mínimos cuadrados y máxima verosimilitud, es que los mínimos cuadrados se pueden utilizar sin hacer suposiciones acerca de las distribuciones de las variables respuestas $Y_i^{'}s$. En cambio para obtener estimadores de máxima verosimilitud es necesario especificar la distribución de probabilidad conjunta de las $Y_i^{'}s$. Sin embargo, para obtener la distribución de muestreo de estimadores de mínimos cuadrados $\boldsymbol{b}$ se requiere supuestos adicionales sobre los $Y_i^{'}s$. Por tanto en la práctica hay pocas ventajas en el uso de los mínimos cuadrados a menos que las ecuaciones de estimación sean computacionalmente más simples.

2.3.3. Estimación de parámetros de modelos lineales generalizados

Deseamos obtener los estimadores de máxima-verosimilitud de los parámetros $\boldsymbol{\beta}$ de los modelos lineales generalizados definidos en la sección [2.2].
La función log-verosimilitud para las respuestas independientes $Y_1, \cdots, Y_N$, que tienen función de densidad conjunta

$$f(y_1, \cdots, y_N; \theta_1, \cdots, \theta_N) = \exp\left[\sum_{i=1}^{N} y_i b(\theta_i) + \sum_{i=1}^{N} c(\theta_i) + \sum_{i=1}^{N} d(y_i)\right]$$

Puede ser escrita como

$$l(\boldsymbol{\theta; y}) = \log \boldsymbol{L(\theta; y)} = \sum y_i b(\theta_i) + \sum c(\theta_i) + \sum d(y_i)$$

con

$$E(Y_i) = \mu_i = -c'(\theta_i)/b'(\theta_i) \tag{2.8}$$

y

$$g(\mu_i) = \boldsymbol{x_i^T \beta} = \sum_{j=1}^{p} x_{ij}\beta_j = \eta_i \tag{2.9}$$

Donde g es una función monótona y diferenciable. También

$$Var(Y_i) = \left[b''(\theta_i)c'(\theta_i) - c''(\theta_i)b'(\theta_i)\right] / \left[b'(\theta_i)\right]^3 \tag{2.10}$$

A continuación se justifican las ecuaciones [2.8] y [2.10].

La distribución de cada Y_i es de la forma canónica y depende de un solo parámetro θ_i, así

$$f(y_i; \theta_i) = \exp\left[y_i b(\theta_i) + c(\theta_i) + d(y_i)\right]$$
$$l_i(\theta_i; y_i) = \log f(y_i; \theta_i)$$

Esto es,

$$l_i = y_i b(\theta_i) + c(\theta_i) + d(y_i) \tag{2.11}$$

Muchos de los resultados clave acerca de los modelos lineales generalizados se refieren a la derivada

$$U_i = \frac{dl_i}{d\theta_i} = y_i b'(\theta_i) + c'(\theta_i)$$

que se denomina la puntuación (score). Para encontrar los momentos de U_i usamos la identidad

$$\frac{d\log f(y_i;\theta_i)}{d\theta_i} = \frac{1}{f(y_i;\theta_i)}\frac{df(y_i;\theta_i)}{d\theta_i} \qquad (I)$$

Por tanto

$$E(U_i) = \int \frac{d\log f(y_i;\theta_i)}{d\theta_i} f(y_i;\theta_i)dy_i = \int \frac{df(y_i;\theta_i)}{d\theta_i}dy_i$$

Donde la integral se toma en el dominio de y_i. Bajo ciertas condiciones de regularidad el término de la derecha es

$$\int \frac{df(y_i;\theta_i)}{d\theta_i}dy_i = \frac{d}{d\theta_i}\int f(y_i;\theta_i)dy_i = \frac{d}{d\theta_i}1 = 0$$

Por tanto

$$E(U_i) = 0$$

Esto es

$$E(U_i) = E(y_i b'(\theta_i) + c'(\theta_i)) = 0$$

De donde

$$E(Y_i) = \mu_i = -c'(\theta_i)/b'(\theta_i)$$

Por otro lado, si diferenciamos respecto a θ_i la identidad considerada (I) y se dan las condiciones para que el orden de las operaciones se puedan intercambiar, tenemos

$$\frac{d}{d\theta_i}\int \frac{d\log f(y_i;\theta_i)}{d\theta_i} f(y_i;\theta_i)dy_i = \frac{d^2}{d\theta_i^2}\int f(y_i;\theta_i)dy_i$$

El lado derecho de la ecuación es cero puesto que $\int f(y_i;\theta_i)dy_i = 1$ y el lado izquierdo puede ser expresado como

$$\int \frac{d^2\log f(y_i;\theta_i)}{d\theta_i^2} f(y_i;\theta_i)dy_i + \int \frac{d\log f(y_i;\theta_i)}{d\theta_i}\frac{df(y_i;\theta_i)}{d\theta_i}dy_i$$

Por tanto, sustituyendo la identidad (I) en el segundo término, obtenemos

$$\int \frac{d^2\log f(y_i;\theta_i)}{d\theta_i^2} f(y_i;\theta_i)dy_i + \int \left[\frac{d\log f(y_i;\theta_i)}{d\theta_i}\right]^2 f(y_i;\theta_i)dy_i = 0$$

Luego

$$E\left[-\frac{d^2\log f(y_i;\theta_i)}{d\theta_i^2}\right] = E\left\{\left[\frac{d\log f(y_i;\theta_i)}{d\theta_i}\right]^2\right\}$$

En términos de U_i tenemos

$$E(-U_i') = E(U_i^2)$$

Donde U_i' denota la derivada de U_i con respecto a θ_i. Como $E(U_i) = 0$, la varianza de U_i, la cual es llamada la **información**, es

$$Var(U_i) = E(U_i^2) = E(-U_i')$$

Por tanto

$$Var(y_i b'(\theta_i) + c'(\theta_i)) = E(-y_i b''(\theta_i) - c''(\theta_i))$$

Esto es

$$[b'(\theta_i)]^2 Var(Y_i) = \{-b''(\theta_i)E(Y_i) - c''(\theta_i)\}$$

Como

$$E(Y_i) = -c'(\theta_i)/b'(\theta_i)$$

se concluye que

$$Var(Y_i) = \left[b''(\theta_i)c'(\theta_i) - c''(\theta_i)b'(\theta_i)\right] / \left[b'(\theta_i)\right]^3$$

En general si consideramos variables aleatorias independientes $Y_1, \cdots, Y_N$ que tienen función de densidad conjunta pertenecientes a la familia exponencial. Una propiedad de la familia de distribuciones exponenciales es que cumplen las condiciones suficientes de regularidad para asegurar que el máximo global de la función logarítmica de verosimilitud $l(\boldsymbol{\theta};\boldsymbol{y})$ viene dada únicamente por la solución de las ecuaciones $\dfrac{\partial l}{\partial \boldsymbol{\theta}} = 0$, o equivalentemente por la solución de $\dfrac{\partial l}{\partial \boldsymbol{\beta}} = 0$ (ver Cox y Hinkley, 1974, Ch. 9).
La derivada parcial de $l(\boldsymbol{\theta};\boldsymbol{y})$ respeto a β_j se define como:

$$U_j = \frac{\partial l(\boldsymbol{\theta};\boldsymbol{y})}{\partial \beta_j} = \sum_{i=1}^{N}\frac{\partial l_i}{\partial \beta_j}$$

Donde

$$l_i = y_i b(\theta_i) + c(\theta_i) + d(y_i)$$

Probemos que

$$U_j = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)$$

Donde x_{ij} es el $j-$ésimo elemento de $\boldsymbol{x_i^T}$.
Para obtener U_j usamos

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j}$$

Diferenciando [2.11] y utilizando [2.8] obtenemos

$$\frac{\partial l_i}{\partial \theta_i} = y_i b'(\theta_i) + c'(\theta_i) = b'(\theta_i)(y_i - \mu_i)$$

Diferenciando [2.8] y utilizando [2.10] tenemos

$$\frac{\partial \mu_i}{\partial \theta_i} = -\frac{c''(\theta_i)}{b'(\theta_i)} + \frac{c'(\theta_i)b''(\theta_i)}{[b'(\theta_i)]^2} = b'(\theta_i)Var(Y_i)$$

Diferenciando [2.9]

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}\frac{\partial \mu_i}{\partial \eta_i}$$

Por lo tanto

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i}\frac{\partial \mu_i}{\partial \beta_j}\Big/\frac{\partial \mu_i}{\partial \theta_i} = \frac{(y_i - \mu_i)x_{ij}}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right) \tag{2.12}$$

De donde

$$U_j = \sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right) \tag{2.13}$$

Como se quería demostrar, donde x_{ij} es el $j-$ésimo elemento de $\boldsymbol{x_i^T}$.
La **matriz de información** se define como la matriz de varianza-covarianza de los U_js, $\boldsymbol{I} = E(\boldsymbol{U}\boldsymbol{U}^T)$ donde $\boldsymbol{U}$ es el vector $U_1, \cdots, U_p$. Por tanto los elementos de la matriz de información se definen por

$$I_{jk} = E\left[U_j U_k\right] = E\left[\frac{\partial l}{\partial \beta_j}\frac{\partial l}{\partial \beta_k}\right]$$

De [2.12] para cada Y_i, la contribución para I_{ij} es

$$E\left[\frac{\partial l_i}{\partial \beta_j}\frac{\partial l_i}{\partial \beta_k}\right] = E\left[\frac{(y_i - \mu_i)^2 x_{ij}x_{ik}}{\{Var(Y_i)\}^2}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right] = \frac{x_{ij}x_{ik}}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

y por tanto

$$I_{ik} = \sum_{i=1}^{N} \frac{x_{ij}x_{ik}}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \tag{2.14}$$

Si Y es una variable aleatoria con función de densidad de probabilidad $f(y;\theta)$ que depende de un solo parámetro θ (o si Y es discreta $f(y;\theta)$ es la función de distribución de probabilidad) se demuestra (ver justificaciones de las ecuaciones [2.8] y [2.10]) que:

$$E(-U') = E(U^2)$$

Donde

$$U = \frac{dl(\theta;y)}{d\theta}, \qquad l(\theta;y) = \log f(y;\theta)$$

U' denota la derivada de U con respecto a θ. Puesto que $E(U) = 0$, la varianza de U la cual es llamada la **información**, es

$$VarU = E(U^2) = E(-U')$$

Por un argumento análogo, se puede demostrar que

$$E\left[\frac{\partial l_i}{\partial \beta_j}\frac{\partial l_i}{\partial \beta_k}\right] = E\left[-\frac{\partial^2 l_i}{\partial \beta_j \beta_k}\right]$$

Por tanto, los elementos de la matriz de información también se dan por

$$I_{jk} = E\left[\frac{\partial l}{\partial \beta_j}\frac{\partial l}{\partial \beta_k}\right] = E\left[-\frac{\partial^2 l}{\partial \beta_j \beta_k}\right] \tag{2.15}$$

Como se ha dicho el máximo global de la función logarítmica de verosimilitud $l(\boldsymbol{\theta}; \boldsymbol{y})$ viene dada únicamente por la solución de las ecuaciones $\dfrac{\partial l}{\partial \boldsymbol{\theta}} = 0$, o equivalentemente por la solución de $\dfrac{\partial l}{\partial \boldsymbol{\beta}} = 0$. Lo que implica resolver las ecuaciones $U_j = 0$ $(j = 1, 2, \cdots, p)$.

En general las ecuaciones $U_j = 0$ $(j = 1, 2, \cdots, p)$ son no lineales y tienen que ser resueltas por iteración numérica. Si el **método Newton-Raphson** es utilizado entonces la $m-$ ésima aproximación está dada por

$$\boldsymbol{b}^{(m)} = \boldsymbol{b}^{(m-1)} - \left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}\right]^{-1}_{\boldsymbol{\beta}=\boldsymbol{b}^{(m-1)}} \boldsymbol{U}^{(m-1)} \tag{2.16}$$

Donde

$$\left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}\right]_{\boldsymbol{\beta}=\boldsymbol{b}^{(m-1)}}$$

Es la matriz de segundas derivadas de l evaluada en $\boldsymbol{\beta} = \boldsymbol{b}^{(m-1)}$ y $\boldsymbol{U}^{(m-1)}$ es el vector de primeras derivadas $U_j = \partial l / \partial \beta_j$ evaluadas en $\boldsymbol{\beta} = \boldsymbol{b}^{(m-1)}$ (esta es la versión multidimensional del método Newton-Raphson para encontrar una solución de una ecuación de una sola variable $f(x) = 0$) es decir

$$x^m = x^{m-1} - \frac{f\left[x^{m-1}\right]}{f'\left[x^{m-1}\right]}$$

Un procedimiento alternativo que a veces es más sencillo que el método de Newton-Raphson se llama el **método de puntuación (scoring).** Se trata de sustituir la matriz de segundas derivadas en [2.16] por la matriz de valores esperados

$$E\left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}\right]$$

De esta manera y utilizando [2.15], la ecuación [2.16] se puede plantear como:

$$\boldsymbol{b}^{(m)} = \boldsymbol{b}^{(m-1)} + \left[\boldsymbol{I}^{m-1}\right]^{-1}_{\boldsymbol{\beta}=\boldsymbol{b}^{(\boldsymbol{m-1})}} \boldsymbol{U}^{(m-1)} \tag{2.17}$$

Donde $\boldsymbol{I}^{(m-1)}$ denota la matriz de información evaluada en $\boldsymbol{b}^{(m-1)}$. Si multiplicamos a izquierda ambos lados de la ecuación [2.17] por $\boldsymbol{I}^{(m-1)}$ obtenemos

$$\boldsymbol{I}^{(m-1)}\boldsymbol{b}^{(m)} = \boldsymbol{I}^{(m-1)}\boldsymbol{b}^{(m-1)} + \boldsymbol{U}^{(m-1)} \tag{2.18}$$

Para modelos lineales generalizados, por [2.14] la entrada (j, k) de la matriz de información $\boldsymbol{I}$ es

$$I_{jk} = \sum_{i=1}^{N} \frac{x_{ij}x_{ik}}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

Por tanto la matriz de información $\boldsymbol{I}$ puede ser escrita como

$$\boldsymbol{I} = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}$$

Donde $\boldsymbol{W}$ es una matriz diagonal de orden $N \times N$ con elementos

$$w_{ii} = \frac{1}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$$

La expresión en el lado derecho de [2.18] es el vector con elementos

$$\sum_k \sum_i \frac{x_{ij} x_{ik}}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 b_k^{(m-1)} + \sum_i \frac{(y_i - \mu_i) x_{ij}}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)$$

Evaluado en $\boldsymbol{b}^{(m-1)}$; esto se desprende de las ecuaciones [2.14] y [2.13]. Por lo tanto el lado derecho de la ecuación [2.18] se puede escribir como

$$\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{z}$$

Donde $\boldsymbol{z}$ tiene elementos

$$z_i = \sum_k x_{ik} b_k^{(m-1)} + (y_i - \mu_i) \left(\frac{\partial \eta_i}{\partial \mu_i} \right) \tag{2.19}$$

Con μ_i y $\partial \eta_i / \partial \mu_i$ evaluados en $\boldsymbol{b}^{(m-1)}$.
Por lo tanto, la ecuación iterativa para el método de puntuación, [2.18] se puede escribir como

$$\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \boldsymbol{b}^{(m)} = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{z} \tag{2.20}$$

Esto tiene la misma forma que las ecuaciones normales para un modelo lineal obtenido por mínimos cuadrados ponderados, excepto que [2.20] tiene que resolverse de forma iterativa porque en general $\boldsymbol{z}$ y $\boldsymbol{W}$ dependen de $\boldsymbol{b}$. Por tanto para modelos lineales generalizados los estimadores de máxima verosimilitud son obtenidos por un procedimiento iterativo de mínimos cuadrados ponderados.
Por lo general un ordenador se necesita para resolver [2.20]. La mayoría de los paquetes estadísticos que incluyen los análisis basados en modelos lineales generalizados tienen programas eficientes para el cálculo de las soluciones. Se comienza mediante una aproximación inicial $\boldsymbol{b}^{(0)}$ para evaluar $\boldsymbol{z}$ y $\boldsymbol{W}$, luego [2.20] es resuelta para obtener $\boldsymbol{b}^{(1)}$ que a su vez se utiliza para obtener mejores aproximaciones de $\boldsymbol{z}$ y $\boldsymbol{W}$ y así sucesivamente hasta que se logre una adecuada convergencia. Cuando la diferencia entre aproximaciones sucesivas $\boldsymbol{b}^{(m)}$ y $\boldsymbol{b}^{(m-1)}$ es suficientemente pequeña $\boldsymbol{b}^{(m)}$ se toma como el estimador de máxima-verosimilitud.

2.4. Modelos de respuesta a dosis

Históricamente uno de los primeros ejemplos o usos de modelos tipo regresión para datos binomiales fue en resultados de bioensayos (Finney 1973). Las respuestas fueron las proporciones o porcentajes de «éxito», por ejemplo, la proporción de animales experimentales muertos por varios niveles de dosis de una sustancia toxica. Tales datos a veces se llaman «respuestas cuánticos». El objetivo es describir la probabilidad de «éxitos» ,π, como función de la dosis, x, por ejemplo $\boldsymbol{g}(\boldsymbol{\pi}) = \beta_1 + \beta_2 x$.
Si la distribución de tolerancia $f(s)$ es la distribución uniforme sobre el intervalo $[c_1, c_2]$

$$f(s) = \begin{cases} \dfrac{1}{c_2 - c_1} & \text{si} \quad c_1 \leq s \leq c_2 \\ 0 \quad \text{si no} \end{cases}$$

Entonces

$$\pi = \int_{c_1}^{x} f(s) ds = \frac{x - c_1}{c_2 - c_1} \quad \text{para} \quad c_1 \leq x \leq c_2$$

(ver fig. 2.1).

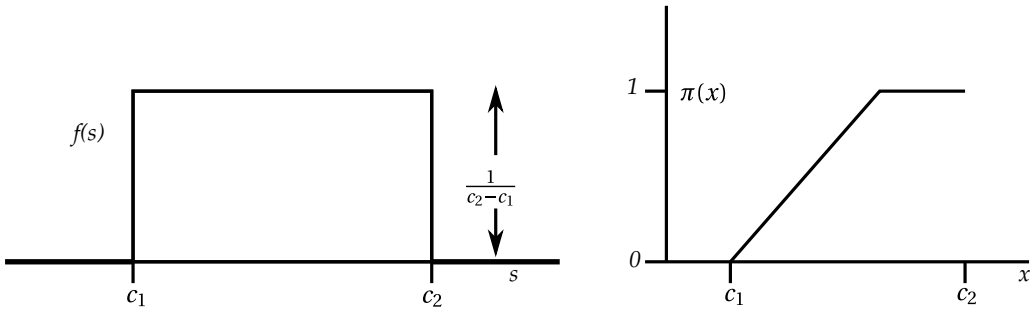

Figura 2.1: Distribución uniforme sobre $[c_1, c_2]$

Esta es la forma

$$\pi = \beta_1 + \beta_2 x \quad \text{donde} \quad \beta_1 = \frac{-c_1}{c_2 - c_1} \quad \text{y} \quad \beta_2 = \frac{1}{c_2 - c_1}$$

Este modelo lineal es equivalente a usar la función identidad como función enlace e imponer condiciones sobre x, β_1 y β_2 correspondientes a $c_1 \leq x \leq c_2$. Estas condiciones extra significan que los métodos estándar para estimar β_1 y β_2 para modelos lineales generalizados no pueden aplicarse directamente. En la práctica este modelo no es ampliamente utilizado.

2.4.1. Modelo Probit

Uno de los modelos originales usados para datos bioexperimentales se llama modelo Probit. La función normal se usa como distribución de tolerancia (ver Figura 2.2)

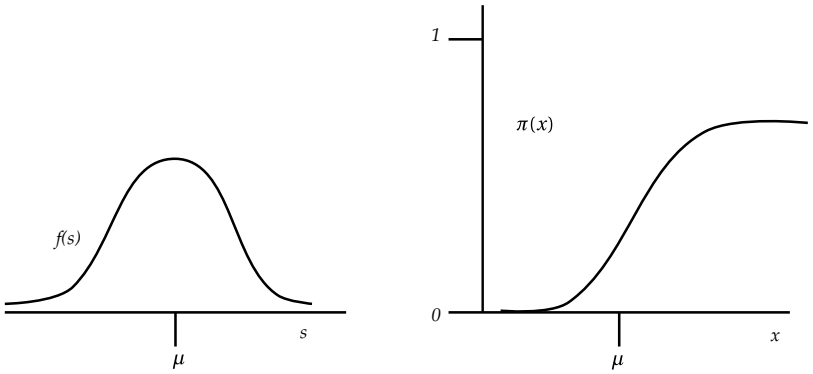

Figura 2.2: Distribución normal $N(\mu, \sigma^2)$

$$\begin{aligned}\pi &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{s-\mu}{\sigma}\right)^2\right] ds \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right)\end{aligned}$$

Donde Φ denota la función de probabilidad acumulada para la distribución normal estándar $N(0,1)$. O sea

$$\Phi^{-1}(\pi) = \beta_1 + \beta_2 x$$

Donde $\beta_1 = -\frac{\mu}{\sigma}$ y $\beta_2 = \frac{1}{\sigma}$ y la función de enlace g es la inversa de la función de probabilidad normal acumulada Φ^{-1}. Estos modelos Probit se usan en varias áreas de ciencias sociales y biológicas en los cuales hay interpretaciones

naturales del modelo; por ejemplo, $x = \mu$ se llama dosis mediana letal $LD(50)$ porque corresponde a la dosis requerida para matar en promedio la mitad de animales.

2.4.2. Modelo Logístico o Modelo Logit

Otro modelo que da resultados numéricos muy parecidos a los del modelo probit, pero que computacionalmente es un poco más fácil es el modelo **logístico** o **modelo logit**. La distribución de tolerancia es:

$$f(s) = \frac{\beta_2 \exp(\beta_1 + \beta_2 s)}{[1 + \exp(\beta_1 + \beta_2 s)]^2}$$

Y así

$$\pi = \int_{-\infty}^{x} f(s)ds = \frac{\exp(\beta_1 + \beta_2 x)}{1 + \exp(\beta_1 + \beta_2 x)} = \{1 + \exp(-\beta - \beta_2 x)\}^{-1}$$

La cual da la función enlace como

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_1 + \beta_2 x$$

$\log\left(\frac{\pi}{1-\pi}\right)$ se llama a veces **función logit** y tiene una interpretación natural como logaritmo de odds. Los odds se definen como la relación π, la probabilidad de éxito $(Y = 1)$ a $1 - \pi$ la probabilidad de no éxito $(Y = 0)$, se simboliza esta relación como

$$\text{odds} = \frac{\pi}{1-\pi} = \frac{\mu}{1-\mu}.$$

Por lo tanto, el modelo de regresión postula una relación entre las probabilidades de registro de la enfermedad y el factor de riesgo. Esto hace que este modelo sea clave y de importancia en epidemiología.

Supóngase que se tiene un valor de probabilidad de 0,5. El odds, según la anterior fórmula es $\frac{0{,}5}{1-0{,}5} = 1$, cuando se reflexiona sobre esto, se ve que tiene perfecto sentido. Si la probabilidad de Y es 0,5, esto significa en términos coloquiales $50 : 50$. Ninguna alternativa se favorece el odds es 1. Un π de 0,2 da un odds de $\frac{0{,}2}{1-0{,}2} = 0{,}25$.

El lado derecho $\beta_1 + \beta_2 x$ del modelo es exactamente como para un modelo de regresión lineal simple y se llama «predictor lineal». Nótese que el modelo,

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_1 + \beta_2 x$$

Debería estrictamente llamarse «regresión logística lineal simple» por lo que hay solo una variable x que se asume tener efectos lineales sobre el logit. El modelo logístico es ampliamente usado para datos binomiales y es implementado en muchos programas estadísticos. Las formas de las funciones $f(s)$ y $\pi(x)$ son similares a las del modelo probit (Figura 2.2) excepto en las colas de las distribuciones.

Los calculos de los odds, para cualquier valor dado de x, salen de evaluar la expresión $\exp(\beta_1 + \beta_2 x)$ con β_1 y β_2 dados por un paquete de computador.

Varios modelos se usan también para datos de respuestas a dosis. Por ejemplo, si la **distribución valor extremo**

$$f(s) = \beta_2 \exp\left[(\beta_1 + \beta_2 s) - \exp(\beta_1 + \beta_2 s)\right]$$

se usa como distribución de tolerancia, entonces

$$\pi = 1 - \exp\left[-\exp(\beta_1 + \beta_2 x)\right]$$

y por lo tanto $\log\left[-\log(1-\pi)\right] = \beta_1 + \beta_2 x$. Esta función enlace, $\log\left[-\log(1-\pi)\right]$ es llamada **función log complementaria**. El modelo es similar a los modelos probit y logísticos para valores de π cerca de 0,5 pero difiere de estos para π cerca a 0 o 1. En el siguiente ejemplo se ilustran estos modelos.

Ejemplo 2.4.1 (Modelos de respuesta a dosis). *El cuadro 2.2 muestra el número de insectos muertos después de una exposición de 5 horas al sulfuro de carbónico gaseoso en varias concentraciones (datos de Bliss 1935). La figura 2.3 muestra las proporciones* $p_i = \frac{y_i}{n_i}$ *graficadas contra la dosis* x_i*. Comenzamos adjuntando el modelo logístico*

$$\pi_i = \frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)}$$

Así

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 x_i$$

y

$$\log(1-\pi_i) = -\log\left[1 + \exp(\beta_1 + \beta_2 x_i)\right]$$

Luego de la ecuación [2.4] la función de log-verosimilitud es:

$$l = \left[\sum_{i=1}^{N} y_i(\beta_1 + \beta_2 x_i) - n_i \log\left[1 + \exp(\beta_1 + \beta_2 x_i)\right] + \log\binom{n_i}{y_i}\right]$$

Y las derivadas respecto a β_1 *y* β_2 *son:*

$$U_1 = \frac{\partial l}{\partial \beta_1} = \sum\left\{y_i - n_i\left[\frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)}\right]\right\} = \sum(y_i - n_i\pi_i)$$

$$U_2 = \frac{\partial l}{\partial \beta_2} = \sum\left\{y_i x_i - n_i x_i\left[\frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)}\right]\right\} = \sum x_i(y_i - n_i\pi_i)$$

Dosis $x_i(\log_{10} CS_2 mgl^{-1})$	Número de insectos, n_i	Número de muertos y_i
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

Cuadro 2.2: Datos de mortalidad de insectos

Similarmente la matriz de información es

$$I = \begin{bmatrix} \sum n_i\pi_i(1-\pi_i) & \sum n_i x_i \pi_i(1-\pi_i) \\ \sum n_i x_i \pi_i(1-\pi_i) & \sum n_i x_i^2 \pi_i(1-\pi_i) \end{bmatrix}$$

Los estimadores de máxima verosimilitud se obtienen de [2.18] resolviendo la ecuación iterativa

$$\boldsymbol{I}^{(m-1)}\boldsymbol{b}^{(m)} = \boldsymbol{I}^{(m-1)}\boldsymbol{b}^{(m-1)} + \boldsymbol{U}^{(m-1)}$$

Donde el superíndice (m) *indica la* $m-$*ésima aproximación y* $\boldsymbol{b}$ *es el vector de estimadores. Partiendo de* $b_1^{(0)} = 0$ *y de* $b_2^{(0)} = 0$ *las aproximaciones sucesivas se muestran en el cuadro 2.3 junto con los valores ajustados* $\widehat{y}_i = n_i\widehat{\pi}_i$*. La matriz de varianza-covarianza estimada para* $\boldsymbol{b}$ *es* $[\boldsymbol{I}(\boldsymbol{b})]^{-1}$*.*
El estadístico de razón de log-verosimilitud es

$$D = \sum_{i=1}^{N}\left[y_i \log\left(\frac{y_i}{\widehat{y}_i}\right) + (n_i - y_i)\log\left(\frac{n-y_i}{n-\widehat{y}_i}\right)\right]$$

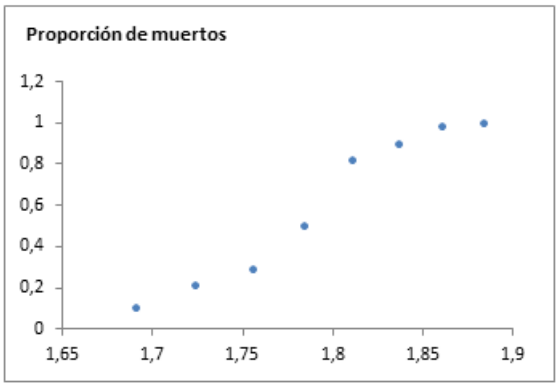

Figura 2.3: Proporción de muertos contra la dosis

		Estimación Inicial	Primera aprox	Segunda aprox	Cuarta aprox	Decima aprox
	b_1	0	-37.849	-53.851	-60.700	-60.717
	b_2	0	21.334	30.382	34.261	34.270
	Observaciones		Valores ajustados			
y_1	6	29.5	8.508	4.544	3.460	3.458
y_2	13	30.0	15.369	11.254	9.845	9.842
y_3	18	31.0	24.810	23.059	22.454	22.451
y_4	28	28.0	30.983	32.946	33.896	33.898
y_5	52	31.5	43.361	48.197	50.092	50.096
y_6	53	29.5	46.739	51.704	53.288	53.291
y_7	61	31.0	53.593	58.060	59.220	59.222
y_8	60	30.0	54.732	58.036	58.742	58.743

$$[\mathbf{I}(\mathbf{b})]^{-1} = \begin{bmatrix} 26.802 & 15.061 \\ 15.061 & 8.469 \end{bmatrix}, \qquad D = 11.23$$

Cuadro 2.3: Ajuste del modelo logístico a los datos de la mortalidad del bicho

Donde $\widehat{y}_i$ denota el valor ajustado.
Los estimadores son $b_1 = -60{,}72$ y $b_2 = 34{,}27$ y sus errores estándar son $\sqrt{26{,}802} = 5{,}18$ y $\sqrt{8{,}469} = 2{,}91$ respectivamente. Si el modelo logístico ofrece un buen resumen de los datos, el estadístico cociente de log-verosimilitud D tiene una distribución χ^2_6 aproximada porque $N = 8$ observaciones y $p = 2$ parámetros. Pero el punto correspondiente al 5 % superior de la distribución χ^2_6 es 12,59 lo cual sugiere que el modelo no se ajusta a los datos particularmente bien.
Usando el programa GLIM, varios modelos alternativos fueron ajustados a estos datos

1. *Logístico (con función enlace logit).*

2. *Probit (con función enlace la función inversa ϕ^{-1} de la Normal acumulada).*

3. *Valor extremo (con la función de enlace* log *complementaria).*
   Los resultados se muestran en el cuadro 2.4. Entre estos modelos el modelo de valor extremo claramente provee la descripción de los datos.

Valor observado de Y	Modelo Logístico	Modelo Probit	Modelo Valor extremo
6	3.46	3.36	5.59
13	9.84	10.72	11.28
18	22.45	23.48	20.95
28	33.90	33.82	30.37
52	50.10	49.62	47.78
53	53.29	53.32	54.14
61	59.22	59.66	61.11
60	58.74	59.23	59.95
D	11.23	10.12	3.45

Cuadro 2.4: Comparación de varios modelos de respuesta-dosis para los datos de mortalidad de bichos

2.5. Regresión logística general

El modelo logístico simple $\log [\pi_i / 1 - \pi_i] = \beta_1 + \beta_2 x_i$ usado en el ejemplo 2.4.1 es un caso especial del modelo de regresión logística general

$$\text{logit}\pi_i = \log \left(\frac{\pi_i}{1 - \pi_i} \right) = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

Donde $\boldsymbol{x_i}$ es un vector de mediciones continuas correspondientes a las covariables y las variables ficticias correspondientes a los niveles de los factores y $\boldsymbol{\beta}$ es el vector de parámetros. Este modelo es ampliamente usado para analizar datos multivariados que involucran respuestas binarias. Provee una técnica poderosa análoga a la regresión múltiple y ANOVA para respuestas continuas. Programas de computador para llevar a cabo regresión logística son disponibles en muchos paquetes estadísticos, por ejemplo el programa PLR en BMDF o el procedimiento PROC LOGIST en SAS. El objetivo de un modelo de regresión logística es entender una respuesta proporcional o binaria (variable dependiente) sobre la base de una o más predicciones.

2.6. Estimación de máxima verosimilitud y el estadístico cociente de log-verosimilitud

Para cualquiera de los modelos de respuesta-dosis y para extensiones tales como los estimadores de máxima verosimilitud del modelo logístico general de los parámetros $\boldsymbol{\beta}$ y consecuentemente de las probabilidades $\pi_i = g^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta})$, se obtienen maximizando la función de log-verosimilitud

$$l(\boldsymbol{\pi; y}) = \sum_{i=1}^{N} \left[y_i \log \pi_i + (n_i - y_i) \log(1 - \pi_i) + \log \binom{n_i}{y_i} \right]$$

Usando los métodos descritos en la sección 2.3. La estimación máxima verosimilitud es posible aun si $n_i = 1$ y/o $y_i = 0$ (a diferencia de algunos de los métodos de mínimos cuadrados) Para medir la bondad de ajuste de un modelo se usa el estadístico de log-verosimilitud

$$D = 2 \left[l(\widehat{\boldsymbol{\pi}}_{\textbf{máx}}; \boldsymbol{y}) - l(\widehat{\boldsymbol{\pi}}; \boldsymbol{y}) \right]$$

Donde $\widehat{\boldsymbol{\pi}}_{\textbf{máx}}$ es el vector de estimadores de máxima verosimilitud correspondiente al modelo maximal y $\widehat{\boldsymbol{\pi}}$ es el vector de estimadores para el modelo de interés.
Sin pérdida de generalidad, para el modelo maximal se toma los $\pi_i's$ como los parámetros a ser estimados. Entonces

$$\frac{\partial l}{\partial \pi_i} = \frac{y_i}{\pi_i} - \frac{n_i - y_i}{1 - \pi_i}$$

Así el $i-$ésimo elemento de $\widehat{\boldsymbol{\pi}}_{\textbf{máx}}$, la solución de la ecuación $\dfrac{\partial l}{\partial \pi_i} = 0$, es y_i / n_i (o sea, la proporción observada de éxitos en el subgrupo i).
Entonces

$$l(\widehat{\boldsymbol{\pi}}_{\textbf{máx}}; \boldsymbol{y}) = \sum_{i=1}^{N} \left[y_i \log \left(\frac{y_i}{n_i} \right) + (n_i - y_i) \log \left(1 - \frac{y_i}{n_i} \right) + \log \binom{n_i}{y_i} \right]$$

Y por tanto

$$D = 2\sum_{i=1}^{N}\left[y_i \log\left(\frac{y_i}{n_i\widehat{\pi}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i\widehat{\pi}_i}\right)\right] \tag{2.21}$$

Luego D es de la forma

$$D = 2\sum O \log\frac{O}{e}$$

Donde O denota las frecuencias observadas y_i y $(n_i - y_i)$ de las celdas del cuadro 2.1 y e denota las frecuencias estimadas correspondientes de valores ajustados $n_i\widehat{\boldsymbol{\pi}_i}$ y $(n_i - n_i\widehat{\pi}_i)$. La suma es sobre todas las $2 \times N$ celdas del Cuadro 2.1.
Nótese que D no involucra ningún parámetro fastidioso (a diferencia de σ^2 para datos de respuesta normal) y así la bondad de ajuste puede asegurarse y las hipótesis pueden ser probadas usando la aproximación

$$D \sim \chi^2_{N-p}$$

Donde p es el número de parámetros β estimados.

2.7. Otros criterios para bondad de ajustes

En vez de una estimación de máxima verosimilitud, se podría estimar los parámetros minimizando la suma ponderada de cuadrados

$$S_W = \sum_{i=1}^{N}\frac{(y_i - n_i\pi_i)^2}{n_i\pi_i(1-\pi_i)}$$

Como $E(Y_i) = n_i\pi_i$ y $Var(Y_i) = n_i\pi_i(1-\pi_i)$.
Esto es equivalente a minimizar el **estadístico chi-cuadrado de Pearson**

$$\boldsymbol{\chi^2 = \sum\frac{(0-e)^2}{e}}$$

Donde 0 representa las frecuencias observadas en el cuadro 2.1, e representa las frecuencias esperadas obtenidas del modelo y la suma es sobre todas las $2 \times N$ celdas de la tabla. La razón es

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{N}\frac{(y_i - n_i\pi_i)^2}{n_i\pi_i} + \sum_{i=1}^{N}\frac{[(n_i - y_i) - n_i(1-\pi_i)]^2}{n_i(1-\pi_i)} \\ &= \sum_{i=1}^{N}\frac{(y_i - n_i\pi_i)^2}{n_i\pi_i(1-\pi_i)}(1 - \pi_i + \pi_i) = S_W\end{aligned}$$

Cuando $\boldsymbol{\chi^2}$ es evaluado en las frecuencias esperadas estimadas, el estadístico es

$$\boldsymbol{\chi^2} = \sum_{i=1}^{N}\frac{(y_i - n_i\widehat{\pi}_i)^2}{n_i\widehat{\pi}_i(1-\widehat{\pi}_i)}$$

El cual es asintóticamente equivalente al estadístico cociente log-verosimilitud en [2.21]

$$D = 2\sum_{i=1}^{N}\left[y_i \log\left(\frac{y_i}{n_i\widehat{\pi}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i\widehat{\pi}_i}\right)\right]$$

Para la demostración se usa la expansión en serie de Taylor de $s\log(s/t)$ al rededor de $s = t$, o sea

$$s\log\frac{s}{t} = (s-1) + \frac{1}{2}\frac{(s-t)^2}{t} + \cdots$$

Así

$$D = 2\sum_{i=1}^{N}\left\{(y_i - n_i\widehat{\pi}_i) + \frac{1}{2}\frac{(y_i - n_i\widehat{\pi}_i)^2}{n_i\widehat{\pi}_i} + [(n_i - y_i) - (n_i - n_i\widehat{\pi}_i)] + \frac{1}{2}\frac{[(n_i - y_i) - (n_i - n_i\widehat{\pi}_i)]^2}{n_i - n_i\widehat{\pi}_i} + \cdots\right\}$$
$$\cong \sum_{i=1}^{N}\frac{(y_i - n_i\widehat{\pi}_i)^2}{n_i\widehat{\pi}_i(1-\widehat{\pi}_i)} = \boldsymbol{X}^2$$

La distribución muestral de D bajo la hipótesis de que el modelo es correcto, es $D \sim \chi^2_{N-p}$, o sea aproximadamente $\boldsymbol{X}^2 \sim \chi^2_{N-p}$.
Otro criterio para bondad de ajuste es el **estadístico chi-cuadrado modificado** obtenido reemplazando las probabilidades estimadas en el denominador de $\boldsymbol{X^2}$ por las frecuencias relativas

$$\boldsymbol{X}^2_{mod} = \sum_{i=1}^{N}\frac{(y_i - n_i\widehat{\pi}_i)^2}{y_i(n_i - y_i)/n_i}$$

Asintóticamente este también tiene distribución χ^2_{N-p} si el modelo es correcto. La escogencia entre $D, \boldsymbol{X}^2$ y $\boldsymbol{X}^2_{mod}$ depende de la adecuación de la aproximación a la distribución χ^2_{N-p}. Hay evidencia que sugiere que $\boldsymbol{X}^2$ es a menudo mejor que D porque D es indebidamente influido por frecuencias muy pequeñas (Cressie y Reaad, 1989). Todas las aproximaciones van a ser pobres si las frecuencias esperadas son demasiado pequeñas (por ejemplo menor que 1).

2.8. Métodos de mínimos cuadrados

Hay algunas ventajas computacionales en el uso de los mínimos cuadrados ponderados en lugar de máxima verosimilitud, en particular si la iteración se puede evitar.
Considere una función ψ de la proporción de éxitos $P_i = Y_i/n_i$, en el $i-$ ésimo subgrupo. La expansión en serie de Taylor de $\psi(P_i)$ alrededor de $P_i = \pi_i$ es:

$$\psi(P_i) = \psi\left(\frac{Y_i}{n_i}\right) = \psi(\pi_i) + \left(\frac{Y_i}{n_i} - \pi_i\right)\psi'(\pi_i) + 0\left(\frac{1}{n_i^2}\right)$$

Así, para una primera aproximación,

$$E\left[\psi(P_i)\right] = \psi(\pi_i)$$

Porque $E\left(\frac{Y_i}{n_i}\right) = \pi_i$. También

$$\begin{aligned} Var[\psi(P_i)] &= E[\psi(P_i) - \psi(\pi_i)]^2 \\ &= [\psi'(\pi_i)]^2 E\left[\frac{Y_i}{n_i} - \pi_i\right]^2 \\ &= [\psi'(\pi_i)]^2\frac{\pi_i(1-\pi_i)}{n_i} \end{aligned}$$

Porque

$$E\left[\frac{Y_i}{n_i} - \pi_i\right]^2 = Var(P_i) = \frac{\pi_i(1-\pi_i)}{n_i}$$

Por lo tanto el criterio de mínimos cuadrados ponderados es

$$\boldsymbol{X}^2 = \sum_{i=1}^{N}\frac{[\psi(y_i/n_i) - \psi(\pi_i)]^2}{[\psi'(\pi_i])^2\pi_i(1-\pi_i)/n_i}$$

Algunas opciones comunes de ψ se resumen en el cuadro 2.5.

$\psi(\pi_i)$	$\boldsymbol{X^2}$
π_i	$\sum \frac{(p_i-\pi_i)^2}{\pi_i(1-\pi_i)/n_i}$
$\text{logit}\pi_i$	$\sum \left[(\text{logit}p_i - \text{logit}\pi_i)^2 \pi_i(1-\pi_i)n_i\right]$
$\arcsin\sqrt{\pi_i}$	$\sum 4n_i\left[\arcsin\sqrt{p_i} - \arcsin\sqrt{\pi_i}\right]^2$

Cuadro 2.5: Algunos modelos de mínimos cuadrados ponderados para datos binarios

Primero, si $\psi(\pi_i) = \pi_i$ y $\pi_i = \boldsymbol{x_i^T\beta}$ el criterio $\boldsymbol{X^2}$ modificado es

$$\boldsymbol{X^2_{mod}} = \sum_{i=1}^{N} \frac{(p_i - \boldsymbol{x_i^T\beta})^2}{p_i(1-p_i)/n_i} \tag{2.22}$$

El cual es lineal en $\boldsymbol{\beta}$ y así la estimación no involucra ninguna iteración. Sin embargo, el estimado $\widehat{\pi}_i = \boldsymbol{x_i^T b}$ puede no caer entre 0 y 1.
Segundo, si $\psi(\pi_i) = \text{logit}\pi_i$ y por tanto $\pi_i = \exp(\boldsymbol{x_i^T\beta})/\left[1+\exp(\boldsymbol{x_i^T\beta})\right]$
Entonces

$$\boldsymbol{X^2_{mod}} = \sum_{i=1}^{N} (z_i - \boldsymbol{x_i^T\beta})^2 \frac{y_i(n_i-y_i)}{n_i} \tag{2.23}$$

Donde

$$z_i = \text{logit}p_i = \log\left(\frac{y_i}{n_i - y_i}\right)$$

Este tampoco, involucra iteración y da estimados de los $\pi_i's$ en el rango $[0,1]$. Cox (1970) llama esta la **transformación logística empírica** y recomienda el uso de

$$z_i = \log\left(\frac{y_i + 1/2}{n_i - Y_i + 1/2}\right)$$

En lugar de

$$z_i = \log\left(\frac{Y_i}{n_i - Y_i}\right)$$

Para reducir el sesgo $E(z_i - \boldsymbol{x_i^T\beta})$. El valor mínimo de [2.23] se llama **el estadístico chi-cuadrado logit mínimo** (Berkson, 1953).
Tercero, la transformación, $\psi(\pi_i) = \arcsin\sqrt{\pi_i}$ (con cualquier escogencia de π_i) dícese poseer la **propiedad estabilizadora de varianza**, ya que

$$var[\psi(P_i)] = [\psi'(\pi_i)]^2 \pi_i(1-\pi_i)/n_i = (4n_i)^{-1}$$

Así, el peso no depende de los parámetros o las respuestas y los cálculos son simples y pueden llevarse a cabo en una calculadora normal.

Observación 2.8.1. *Muchos de los asuntos que se presentan en el uso de regresión múltiple para variable de respuesta continua son también relevantes con respuestas binarias. Los test para la inclusión o exclusión de ciertos términos usualmente no son independientes y es necesario establecer cuidadosamente cuales términos se incluyen en el modelo en cada etapa. Si hay muchas variables explicativas (o exploratorias), los métodos de selección en cada paso pueden usarse para identificar los mejores subconjuntos de variables.*
El examen gráfico de los residuales es útil para asegurar la adecuación del modelo propuesto. Una simple definición de residuales estandarizados es

$$r_i = \frac{p_i - \widehat{\pi}_i}{\sqrt{[\widehat{\pi}_i(1-\widehat{\pi}_i)/n_i]}}$$

Donde $p_i = y_i/n_i$ es la proporción observada y $\widehat{\pi}_i$ es la proporción estimada bajo el modelo. Los $\widehat{\pi}_i$'s aproximadamente tienen media zero y desviación estándar uno. Ellos son las raíces cuadradas con signo de las contribuciones al estadístico $\boldsymbol{X}^2$. Cuando se grafican con niveles de factor y covariados no deben mostrar ningún patrón sistémico. Sin embargo sus distribuciones de probabilidad pueden estar lejos de la normal. Residuales mas complicados, los cuales son casi normales se describen en Cox y Snell (1968).

Mas recientemente Pierce y Schafer (1986) han mostrado que las raíces cuadradas signadas de las contribuciones al estadístico D,

$$d_i = \pm\sqrt{2}\left[y_i \log\left(\frac{y_i}{n_i\widehat{\pi}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i\widehat{\pi}_i}\right)\right]$$

son aproximadamente normalmente distribuidas si el modelo es bueno y así proveer residuales apropiados para propósitos de diagnósticos.

Capítulo 3

Conceptos epidemiológicos y estadísticos

3.1. Riesgo y riesgo relativo

En epidemiología, a menudo interesa evaluar el chance de que un individuo que posea cierto atributo sufra de una enfermedad especifica. La medida epidemiológica más básica es la probabilidad de que un individuo se convierta en un nuevo enfermo, dado que el individuo posea el atributo particular bajo consideración. Esto se llama el riesgo de la enfermedad, el atributo considerado se llama factor de riesgo. Luego el riesgo mide la probabilidad de incidencia de la enfermedad.

Aunque el riesgo es un resumen útil de la relación entre factor de riesgo y enfermedad, no es suficiente en si mismo para evaluar la importancia del factor de riesgo al resultado de la enfermedad. Por ejemplo, se puede hallar que el 30 % de una muestra de mujeres que usan un tipo particular de anticonceptivo desarrollan cáncer de seno, y así el riesgo de cáncer de seno es 0,3 para los usuarios de la píldora. Esto parece una evidencia suficiente implicando la píldora, a no ser que se vea claramente que un porcentaje similar de no usuarios también han desarrollado el cáncer de seno. Como en muchos procedimientos en epidemiología, se requiere un grupo de comparación, el más simple de tomar para el ejemplo considerado es el grupo sin el factor de riesgo. Esto lleva a una definición de riesgo relativo (o cociente de riesgo)

Riesgo relativo o cociente de riesgo: es el cociente del riesgo de enfermedad para aquellos con el factor de riesgo al riesgo de enfermedad para aquellos sin el factor de riesgo.

Si el riesgo relativo es mayor que 1, entonces el factor bajo investigación aumenta el riesgo. Si es menor que 1, reduce el riesgo. Un factor con un riesgo relativo menor que 1 se refiere a veces como factor protectivo. En general se usará el termino «factor de riesgo» sin especificar la dirección de su efecto.

El calculo del riesgo y el riesgo relativo es particularmente simple de una tabla de 2×2, estado del factor de riesgo contra estado de enfermo, designados algebraicamente en el Cuadro 3.1. En el cuadro se presentan datos de n sujetos, libres de enfermedad al comienzo del estudio. El riesgo o factor de riesgo de cada individuo se registra como también si el o ella más tarde desarrollaron la enfermedad durante el estudio.

	Estado de la enfermedad		
Estado del factor de riesgo	Enfermo	No enfermo	Total
Expuesto	a	b	$a+b$
No expuesto	c	d	$c+d$
Total	$a+c$	$b+d$	n

Cuadro 3.1: Muestra de datos de un estudio de incidencia

Del cuadro 3.1 se tiene, por ejemplo, que el número de personas con factor de riesgo pero sin la enfermedad es b. En general

$$\text{riesgo=número de casos de enfermedad/número de personas en riesgo} \tag{3.1}$$

Por el cuadro 3.1 los riesgos específicos a exposición, para aquellos con el factor de riesgo son $a/(a+b)$ y para aquellos sin factor de riesgo $c/(c+d)$. El riesgo relativo para aquellos con factor de riesgo, comparado con aquellos

sin tal, está dado por:

$$\frac{a/(a+b)}{c/(c+d)} = \frac{a(c+d)}{c(a+b)} \tag{3.2}$$

En muchas situaciones en la vida real los datos reunidos en un estudio epidemiológico serán una muestra de datos sobre el sujeto de interés. Luego el riesgo y riesgo relativo calculado a partir de los datos son estimaciones para las entidades equivalentes en la población total. Por ejemplo todas, las mujeres pre-menopausicas en el ejemplo anterior relativo al uso de la píldora.
Se debe especificar el error de muestreo inherente a nuestras estimaciones basadas en muestras del valor verdadero del riesgo y el riesgo relativo (poblacional). Como en otros casos, esto se hace mejor especificando el error estándar (variación muestra a muestra en el valor estimado) o un intervalo de confianza. Se usa el símbolo π o R para representar el riesgo de la población y r para representar el riesgo muestral.
Por definición, el riesgo de la población es simplemente una probabilidad y el error estándar es estimado por

$$\widehat{se}(r) = \sqrt{r(1-r)/n} \tag{3.3}$$

Usando una distribución normal de confianza a 95 % para π es:

$$r \pm 1{,}96\widehat{se}(r) \tag{3.4}$$

El intervalo de confianza para el riesgo relativo es ligeramente más difícil de calcular, la distribución del riesgo relativo muestral es cruzada y una transformación *log* es necesaria para asegurar normalidad aproximada. En la escala *log* Katz et al. (1978) mostraron (utilizando la notación del cuadro 3.1) que

$$\widehat{se}(\widehat{\lambda}) = \sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}} \tag{3.5}$$

Donde λ es el riesgo relativo poblacional y $\widehat{\lambda}$ es el estimado en la muestra definido por [3.2] entonces un intervalo de confianza de 95 % para $\log_e \lambda$ es

$$\log_e \widehat{\lambda} \pm 1{,}96\widehat{se}(\log_e \widehat{\lambda})$$

Con límite inferior y superior de

$$L_{\log} = \log_e \widehat{\lambda} - 1{,}96\widehat{se}(\log_e \widehat{\lambda}); \qquad U_{\log} = \log_e \widehat{\lambda} + 1{,}96\widehat{se}(\log_e \widehat{\lambda}) \tag{3.6}$$

Realmente se desea un intervalo de confianza de 95 % para λ mismo. Esto se obtiene elevando los dos límites en [3.6] a la potencia e. Esto es, los límites inferior y superior en el intervalo de confianza de 95 % para el riesgo relativo λ, son:

$$L = \exp(L_{\log}); \qquad U = \exp(U_{\log}) \tag{3.7}$$

Ejemplo 3.1.1. *El proyecto Pooling (Pool 5) estudio factores de riesgo para enfermedad cardiaca coronaria entre hombres en cuatro ciudades de USA Albany, Chicago, Framingham y Tecumseh (Grupo de investigación de proyecto Pooling 1978). El cuadro 3.2 da el status de fumar presente al entrar al estudio, y si o no ocurrió un evento coronario dentro de los siguientes (aproximadamente) 10 años para 1905 hombres en edad de 50-54 al entrar. Cualquier hombre con síntomas coronarios preexistente a la entrada es excluido.*

Por [3.1] el riesgo general de un evento coronario es $216/1905 = 0{,}1134$; el riesgo para fumadores es $166/1342 = 0{,}1237$, mientras que para los no fumadores es $50/563 = 0{,}0888$. Por [3.2], el riesgo relativo es $0{,}1237/00888 = 1{,}393$. Por [3.4], el intervalo de confianza de 95 % para el riesgo de un evento coronario para fumadores es

$$0{,}1237 \pm 1{,}96\sqrt{0{,}1237(1-0{,}1237)/1342}$$

Esto es, $0{,}11237 \pm 0{,}0176$ ó $(0{,}1061, 0{,}1413)$. De [3.5], el error estándar estimado de log del riesgo relativo es

$$\sqrt{\frac{1}{166} - \frac{1}{1342} + \frac{1}{50} - \frac{1}{563}} = 0{,}1533$$

Por [3.6], los límites de confianza de 95 % para el log del riesgo relativo son

$$L_{\log} = \log_e(1{,}393) - 1{,}96 \times 0{,}1533 = 0{,}0310,$$
$$U_{\log} = \log_e(1{,}393) + 1{,}96 \times 0{,}1533 = 0{,}6319$$

	¿Evento coronario durante el seguimiento?		
¿Fumador al entrar?	Si	No	total
Si	166	1176	1342
No	50	513	563
Total	216	1689	1905

Cuadro 3.2: Fumadores y eventos coronarios en el Proyecto Pooling

Por [3.7], los límites de confianza de 95 % para el log del riesgo relativo de un evento coronario para fumadores comparado a no-fumadores son entonces

$$L = \exp(0{,}0310) = 1{,}031$$
$$U = \exp(0{,}6319) = 1{,}881$$

Luego se estima que el riesgo de un evento coronario para fumadores es 0,124, y se está 95 % seguro de que el intervalo $(0{,}106, 0{,}1141)$ contiene el riesgo de la población verdadero. El riesgo relativo estimado de un evento coronario para fumadores comparado a no fumadores es 1,39, y es 95 % seguro que el intervalo $(1{,}03, 1{,}88)$ contiene el riesgo relativo de la población verdadero.

3.1.1. Odds y cocientes de Odds

Como se ha visto, el riesgo es una probabilidad, cuando una probabilidad se calcula es posible calcular una especificación de «chance» llamado los odds. Mientras la probabilidad mide el número de veces que el resultado de interés ocurre (por ejemplo, enfermedad) relativo al número total de observaciones (esto es, el tamaño de la muestra), los odds mide el número de veces que el resultado ocurre relativo al número de veces que no. Los odds pueden ser calculados para grupos diferentes; aquí se estaría interesado en los odds para aquellos expuestos al riesgo o factor de riesgo y los odds para aquellos no expuestos. El cociente de estos dos se llama el cociente de odds (**odds ratio**). Similar al riesgo relativo un cociente de odds mayor que 1 implica que la exposición al factor bajo investigación aumenta los odds de enfermedad, mientras que un valor por debajo de 1 significa que el factor reduce los odds de enfermedad. En general

$$\text{Odds=número de casos de enfermedad/número de casos de no enfermedad} \tag{3.8}$$

Del Cuadro 3.1 los odds específicos-exposición, para aquellos con el factor de riesgo, a/b; y para aquellos sin el factor de riesgo, c/d. El cociente de odds para aquellos con el factor de riesgo, comparado a aquellos sin el factor, esta dado por

$$\widehat{\psi} = \frac{a/b}{c/d} = \frac{ad}{bc} \tag{3.9}$$

La letra griega ψ(psi) es comúnmente usada para representar el cociente de odds de la población, con el sombrero como en [3.9] denota el valor muestral que es un estimado de su población equivalente.

Otro modo de derivar los odds es como el cociente de riesgo a su complemento. Esto puede justificarse fácilmente de [3.1] y [3.8], por ejemplo, los odds para aquellos con el factor de riesgo son dados por $r/(1-r)$.
En la vida cotidiana, se habla de odds muy a menudo en situaciones de apuestas, tales como los odds de un caballo ganando una carrera. Los epidemiólogos toman «odds» y «cocientes de odds» para referirse al chance de una incidencia de la enfermedad, justo como con «riesgo» y «riesgo relativo».

Como con el riesgo relativo, la distribución del cociente de odds es mejor aproximada por una distribuciones normal si se aplica una transformación log. Woolf (1955) mostró que

$$\widehat{se}\left(\log_e \widehat{\psi}\right) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \tag{3.10}$$

Y los límites de confianza de 95 % para ψ son

$$L_{\log} = \log_e \widehat{\psi} - 1{,}96\widehat{se}(\log_e \widehat{\psi}) \qquad U_{\log} = \log_e \widehat{\psi} + 1{,}96\widehat{se}(\log_e \widehat{\psi}) \tag{3.11}$$

Y los límites de confianza de 95 %, para ψ son

$$L = \exp(L_{\log}) \qquad U = \exp(U_{\log}) \tag{3.12}$$

Ejemplo 3.1.2. *Para el proyecto Pooling, del cuadro 3.2 se obtienen los siguientes resultados. Por [3.8] los odds globales de un evento coronario son* $216/1689 = 0{,}1279$*; los odds de un evento coronario para fumadores son* $166/1176 = 0{,}1412$*; mientras que para no fumadores son* $50/513 = 0{,}0975$*. Por [3.9] el cociente de odds para un evento coronario, comparando fumadores a no fumadores es*

$$\frac{166/1176}{50/513} = \frac{0{,}1412}{0{,}0975} = 1{,}448$$

Y por [3.10] el error estimado del log *del cociente Odds es:*

$$\sqrt{\frac{1}{166} + \frac{1}{1176} + \frac{1}{50} + \frac{1}{513}} = 1{,}1698$$

Por [3.11], los límites de confianza de 95 % *para el* log *de los cocientes de Odds son entonces*

$$L_{\log} = \log_e(1{,}448) - 1{,}96 \times 0{,}1698 = 0{,}0374$$
$$U_{\log} = \log_e(1{,}448) + 1{,}96 \times 0{,}1698 = 0{,}7030$$

Por [3.12], los límites de confianza del 95 % *para el cociente odds para un evento coronario comparando fumadores a no fumadores son entonces*

$$L = \exp(0{,}0374) = 1{,}038$$
$$U = \exp(0{,}7030) = 2{,}020$$

Luego, los odds de un evento coronario se estiman como 1,45 *veces para fumadores sobre no fumadores, y se está* 95 % *seguro de que el intervalo* $(1{,}04, 2{,}02)$ *contiene el cociente de odds verdadero.*

3.2. Como decidir sobre una medida de chance comparativo

3.2.1. ¿Riesgo relativo o cocientes de Odds?

Como se ha visto, ambos el riesgo y los odds miden el chance de incidencia de enfermedad de algún modo, y así el riesgo relativo y el cociente de odds miden chance comparativo. El riesgo es la medida preferida porque es una probabilidad y las probabilidades son bien entendidas como versiones a largo plazo de proporciones. Los odds no son tan bien entendidos, y en consecuencia rara vez usados.
La razón de porqué considerar odds, es que el cociente de odds es a menudo una buena aproximación al riesgo relativo, y en algunos casos el cociente es todo lo que se puede estimar (por ejemplo, la situación en estudios de control de casos) o es más conveniente para calcular (en análisis de regresión logística).
El cociente de odds será una buena aproximación al riesgo relativo siempre y cuando la enfermedad en cuestión sea rara. De acuerdo al cuadro 3.1 cuando la enfermedad es rara debe ser que

$$\begin{aligned} a + b &\approx b \\ c + d &\approx d \end{aligned} \tag{3.13}$$

Donde $\approx$ significa «aproximadamente igual a». Así, de [3.2] [3.9] y [3.13]

$$\widehat{\psi} = \frac{ad}{bc} \approx \frac{a(c+d)}{(a+b)c} = \widehat{\lambda}$$

Esto es el cociente de odds muestral $\widehat{\psi}$ (o OR) es aproximadamente al riesgo relativo muestral $\widehat{\lambda}$ (o RR) cuando $a+b \approx b$ y $c+d \cong d$. Los ejemplos 3.1.1 y 3.1.2 muestran que el riesgo relativo (1.59) y el cociente de odds (1.45) para un evento coronario, comparado fumadores a no fumadores son muy similares para el proyecto Pooling.

Epidemiólogos y estadísticos tradicionalmente están de acuerdo que los cocientes de odds es una buena aproximación de los riesgos relativos, cuando la incidencia es relativamente baja (menos del 10 por ciento) entre ambos aquellos con o sin el factor de riesgo. Por ejemplo si pocos no fumadores tienen un evento coronario, pero virtualmente todos los fumadores si, entonces el cociente de odds será sustancialmente mayor que el riesgo relativo.

Ejemplo 3.2.1. *se consideran los resultados hipotéticos del cuadro 3.3. Una columna extra de riesgos se ha agregado para mejorar la interpretación. Esto es a menudo útil en reportes escritos. En este caso la enfermedad es rara globalmente (solo una persona en 100 la tiene) y sin embargo el riesgo relativo es:*

$$\widehat{\lambda} = \frac{0{,}5}{0{,}00102} = 490$$

Mientras que el cociente de odds es:

$$\widehat{\psi} = \frac{9 \times 981}{1 \times 9} = 981$$

Casi el doble. Así, el cociente de odds no da una buena aproximación al riesgo relativo en este caso. El cuadro 3.3 muestra un ejemplo muy extremo de como los cocientes odds pueden sobre estimar el riesgo relativo y se podría argumentar que los valores exactos importan poco cuando la dirección del efecto es obvia. Sin embargo como el riesgo relativo y el cociente de odds no son lo mismo, es apropiado especificar cual de los dos está siendo reportado. Muchas publicaciones epidemiológicas han cometido un error a este respecto usualmente llamando erróneamente un cociente de odds a un riesgo relativo. Seria posible calcular ambos en cualquier investigación a seguir como en los ejemplos 3.1 y 3.2, aunque no en un estudio de control de caso. Sin embargo, normalmente solo se reporta uno.

	Estado de la enfermedad			
Estado del facto de riesgo	Enfermo	No enfermo	Total	Riesgo
Expuesto	9	9	18	0.5
No expuesto	1	981	982	0.00102
Total	10	990	1000	0.01010

Cuadro 3.3: Datos hipotéticos de un estudio de incidencia

3.2.2. Medidas de diferencia

Ambas el riesgo relativo y el cociente de odds, son medidas proporcionales. Tales medidas parecen muy naturales cuando el chance esta siendo medido ya que el chance mismo es medido proporcionalmente. Sin embargo, las medidas de diferencia también pueden ser usadas esto es, la diferencia entre los riesgos o entre los odds. Es generalmente la diferencia de riesgo la que se usa cuando se escoge una medida de diferencia. Así del ejemplo 3.1 la

diferencia de riesgo coronario de fumadores comparado a no fumadores es 01237 − 0,0888 = 0,0349. El riesgo de un evento coronario es alrededor de 0,03 mayor para los fumadores. Esto es discutiblemente no tan fácil de interpretar como la afirmación anterior, usando el riesgo relativo, que los fumadores tienen 1,39 veces el riesgo de un evento coronario. Una intervalo de confianza de 95 % para la diferencia de riesgo puede ser calculado usando

$$p_1 - p_2 \pm 1{,}96\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

3.3. Estudios de prevalencia

En las secciones anteriores se ha definido riesgo, riesgo relativo, odds, cociente de odds y diferencia de riesgo como medidas de chance en estudios de incidencia. Técnicamente no hay razón de porque las mismas definiciones se podrían usar en estudios de prevalencia: las ecuaciones [3.1]-[3.13] podrían todos definirse propiamente en sentido matemático. A diferencia del uso inglés general , los epidemiólogos reservan todos estos términos clave para referirse a incidencia. Sin embargo, esto es relativamente algo desafortunado, ya que es perfectamente natural querer hablar sobre «riesgo» en relación con los datos de prevalencia.

Ejemplo 3.3.1. *Smith y otros (1991) dan los resultados de un estudio transversal de enfermedad vascular periférica (PVD) en Escocia.*
El cuadro 3.4 muestra los casos prevalentes y no-casos de la enfermedad clasificada por el hábito de fumar cigarrillos para los hombres. Por lo tanto:

$$\begin{aligned}
\textit{Riesgo de prevalencia global de PVD} &= 56/4916 = 0{,}0113;\\
\textit{Riesgo de prevalencia de PVD para fumadores} &= 15/1727 = 0{,}00869;\\
\textit{Riesgo de prevalencia de PVD para no fumadores} &= 41/3229 = 0{,}0127;\\
\textit{Riesgo relativo de prevalencia} &= 0{,}00869/0{,}0127 = 0{,}68;\\
\textit{odds de prevalencia de PVD para fumadores} &= 15/1712 = 0{,}00876;\\
\textit{odds de prevalencia de PVD para no fumadores} &= 41/3188 = 0{,}0129\\
\textit{Cociente de odds de prevalencia} &= 0{,}00876/0{,}0129 = 0{,}68
\end{aligned}$$

	¿Enfermedad vascular periférica?		
Fumador de cigarrillo?	Si	No	total
Si	15	1712	1727
No	41	3188	3229
Total	56	4900	4956

Cuadro 3.4: Fumar cigarrillo y enfermedad vascular periférica de hombres Escoceses.

Se observa que el riesgo relativo de prevalencia y el cociente de odds son iguales (a dos cifras decimales) como sería de esperar para una enfermedad tan rara como enfermedad vascular periférica.

Los resultados del ejemplo 3.3.1 parecen ir contra lo esperado: fumar cigarrillo parece ser protectivo (riesgo relativo < 1). Esto ilustra una gran desventaja o defecto de los estudios de prevalencia: individuos con enfermedad preexistente pueden haber alterado su estilo de vida, posiblemente a consejo médico, y así ellos ya no están expuestos al factor de riesgo. En el ejemplo un fumador puede suspender una vez que haya experimentado PVD u otros síntomas cardiovasculares.

El artículo original que dio lugar al ejemplo 3.3.1 distinguía entre exfumadores y los que nunca han fumado al considerar el grupo de no fumador. El cuadro 3.5 muestra los resultados cuando se usa esta clasificación más extendida. Como se espera, aquellos que nunca fumaron tienen la prevalencia más baja, exfumadores tienen la más alta, con los fumadores actuales entre los dos.
En otras situaciones puede no ser posible identificar aquellos que fueron expuestos previamente a cualquier factor de riesgo que sea de interés. Los resultados pueden entonces ser sesgados a favor del factor de riesgo, como en el ejemplo 3.3.1.

	¿Enfermedad vascular periférica?			
Fumador de cigarrillo?	Si	No	total	Prevalencia
Fumador actual	15	1712	1727	0.0087
Ex fumador	33	1897	1930	0.0171
Nunca ha fumado	8	1291	1299	0.0062
Total	56	4900	4956	0.0113

Cuadro 3.5: Fumar cigarrillo y enfermedad vascular periférica en una muestra de hombres Escoceses.

Por supuesto, aun en la situación más extendida mostrada en el cuadro 3.5, no hay garantía de que la enfermedad ha causado que la gente deje el vicio, pero no está probado a no ser que se pueda asegurar la sucesión de eventos para cada fumador a presente.
Todos estos problemas muestran que si se desea demostrar causalidad directa del factor de riesgo a la enfermedad, es aconsejable usar datos de incidencia, más que de prevalencia.

3.4. Probando asociación

El riesgo relativo y el cociente de odds son medidas significativas de asociación entre el factor de riesgo y la enfermedad. Ambas miden el chance relativo de enfermedad con el factor de riesgo comparados a los sin enfermedad. Si el factor supuesto de riesgo no tienen efecto, ambos, el riesgo relativo y el cociente de odds deben resultar ser alrededor de uno. En la práctica se toma una muestra de datos y se mide o el riesgo relativo muestral o el cociente de odds. Se podría entonces considerar si este resultado muestral provee evidencia de que el valor de la población equivalente es diferente de 1,0. Si no, entonces se concluiría que no hay evidencia de una asociación entre el supuesto factor de riesgo y la enfermedad en cuestión. Tal resultado pondría en duda la teoría de que el supuesto factor de riesgo causa la enfermedad.
Considerando el cuadro 3.1, es claro que un test de no asociación entre factor de riesgo y enfermedad se logra por un test Chi-cuadrado. Del cuadro 3.1, los valores esperados (E) cuando la hipótesis nula de no-asociación (equivalente a $H_0 : \lambda = 1$ o $H_0 : \psi = 1$) es válida, son

$$\begin{array}{ll} (a+b)(a+c)/n & (a+b)(b+d)/n \\ (c+d)(a+c)/n & (c+d)(b+d)/n \end{array}$$

respectivamente.
Usando el estadístico de prueba chi-cuadrado

$$\sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{3.14}$$

donde O_{ij} es el valor observado de la celda (i, j) y E_{ij} es el valor esperado, y haciendo una mínima manipulación algebraica obtenemos la forma específica del test estadístico chi-cuadrado para una tabla 2×2:

$$\frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} \tag{3.15}$$

Como hay dos filas y dos columnas, los grados de libertad asociados con el estadístico chi-cuadrado son $(2-1)(2-1)=1$. Luego se compara [3.15] con una chi-cuadrado con un grado de libertad χ^2_1.
Considerando los datos del proyecto Pooling del ejemplo 3.1.1, cuadro 3.2 por [3.15], el estadístico chi-cuadrado es

$$\frac{1905(166\times 513-50\times 1176)^2}{1342\times 563\times 216\times 1689}=4{,}80$$

Utilizando una tabla se encuentra que $\chi^2_1=3{,}84$ en el nivel 5 % y de 5,02 en el nivel 2,5 %. Luego el resultado es significativo al nivel 5 %, pero no al nivel más extremo de 2,5 %. El valor p calculado es 0,028. se concluye que hay una evidencia (aunque no particularmente fuerte) de una asociación entre fumar y un evento coronario. El riesgo relativo y el cociente de odds son ambos significativamente distintos de 1, al nivel de significancia de 5 %.
Algo para notar del resultado, es que no se puede usar para concluir que fumar causa enfermedad cardiaca coronaria. Aunque el resultado ciertamente no refuta la idea de una conexión causal, la asociación podría ser resultado de fuerzas externas. Por ejemplo, podría ser que los fumadores son principalmente ya viejos, y son los viejos los que sufren eventos coronarios. El fumar y la enfermedad coronaria son entonces, tal vez, solo asociados en razón de su conexión causal común con la edad.
El test chi-cuadrado es un test bilateral, y así la conclusión tomando los datos del proyecto Pooling sería $\lambda\neq 1$ o $\psi\neq 1$. Para asegurar la dirección de la conexión entre el factor de riesgo y la enfermedad simplemente se mira el estimado de λ o ψ. Por los ejemplos 3.1.1 y 3.1.2 se tiene que los valores estimados $\hat{\lambda}=1{,}39$ y $\hat{\psi}=1{,}45$. son ambos mayores que 1. Luego se puede concluir que los fumadores más probablemente sufrirán un evento coronario, y se tiene evidencia de que la asociación observada entre fumar y enfermedad coronaria no es simplemente debida al azar.
De los resultados de los ejemplos 3.1.1 y 3.1.2 , los intervalos de confianza de 95 % para λ y ψ son $(1{,}03, 1{,}88)$ y $(1{,}04, 2{,}02)$ respectivamente. Ambos excluyen la unidad, lo cual coincide con el resultado $\lambda\neq 1$ o $\psi\neq 1$ a nivel 5 % de significancia. Los limites inferiores de confianza de 95 % son ambos muy cerca de 1, lo cual se refleja en el valor p al ser solo justo menor de 0,05.

Capítulo 4

Interpretación de los coeficientes de regresión logística

4.1. Introducción

En este trabajo se usan dos medidas estadísticas para explorar asociaciones: odds ratios ajustados (ORs) y riesgos atribuibles a población (PAR) calculados con estos ORs (Bolton y Robinson 2010; Greenland and Drescher 1993). Los ORs es una medida del poder de asociación entre dos variables dicótomas: A más diferente de 1 un ORs es más fuerte esa asociación. Cuando los intervalos de confianza a 95 % (CIs) de dos ORs no se interceptan se consideran significativamente diferentes. Un método multivariado, regresión logística, permite ajustar ORs, para variables potencialmente confusas.
Una vez que se ajustan ORs, si se asume que OR mide una relación causal entre una condición y un factor de riesgo no explicado por factores de confusión, los estimados PAR estiman la reducción porcentual en la prevalencia de la condición que sería observada si el factor de riesgo fuera removido de la población.

4.2. Regresión logística para ajustar Odds y Odds Ratio

Consideremos un modelo logístico simple

$$\begin{aligned} \text{logit}(\pi) &= \log\left[\frac{\pi}{1-\pi}\right] = \beta_1 + \beta_2 x \\ &= \pi = \{1 + \exp(-\beta_1 - \beta_2 x)\} \end{aligned} \tag{4.1}$$

El objetivo de un modelo de regresión logístico es entender una respuesta proporcional o binaria (variable dependiente) sobre la base de una o más predicciones.
Hay dos usos importantes empleados por los estadísticos en un modelo logístico, usados para explorar asociaciones. Uno es la interpretación de los estimados de los parámetros como cocientes de odds (Odds Ratio) o simplemente ORs. El otro esta relacionado al cálculo de los riesgos ajustados, lo cual puede entenderse como la probabilidad de 1 (éxito) o riesgos atribuibles a población (PAR) calculados con estos ORs. Ambos usos juegan papeles importantes en dominios tales como investigación en salud y medicina, crédito, investigación en ciencias sociales, y otros.
Los **Odds** $\left(\frac{\pi}{1-\pi}\right)$ para un valor específico de la variable x, de acuerdo con [4.1], salen de calcular $e^{\beta_1+\beta_2 x}$. De forma análoga si se deseara estimar los **odds ratio**, ψ, para $x = x_1$ comparado a $x = x_0$, obviamente esto se podría hacer calculando los dos separados y dividiendo; un método más rápido es usar el hecho de que el logaritmo de un cociente

es la diferencia de los logaritmos del numerador y denominador

$$\begin{aligned}\log(\psi) &= \log\left(\frac{\pi_1/1-\pi_1}{\pi_0/1-\pi_0}\right) = \log\left(\frac{\text{odds}_1}{\text{odds}_0}\right) = \log(\text{odds}_1) - \log(\text{odds}_0)\\ &= \text{logit}(\pi_1) - \text{logit}(\pi_0)\\ &= (\beta_1 + \beta_2 x_1) - (\beta_1 + \beta_2 x_0)\\ &= \beta_2(x_1 - x_0)\end{aligned}$$

Luego

$$\psi = \exp\{\beta_2(x_1 - x_0)\} \tag{4.2}$$

Esto es el **odds ratio** (x_1 comparado a x_0) se calcula por exponenciación del parámetro (β_2) por la diferencia entre x_1 y x_0. Su error estándar es

$$se(\log\psi) = (x_1 - x_0)se(\beta_2) \tag{4.3}$$

Luego el intervalo de confianza para ψ de 95 % tiene límites

$$\begin{aligned}&\exp\{\beta_2(x_1 - x_0) - 1{,}96(x_1 - x_0)\widehat{se}(\beta_2)\}\\ &\exp\{\beta_2(x_1 - x_0) + 1{,}96(x_1 - x_0)\widehat{se}(\beta_2)\}\end{aligned} \tag{4.4}$$

errores estándar, y por lo tanto intervalos de confianza son más difíciles de obtener para riesgos y riesgos relativos. Muchos estadísticos tienden a interpretar cocientes de odds (odds ratios) como aproximaciones de cocientes de riesgos relativos (risk ratios). Se define un factor de riesgo de valor binario, o predictor de interés como x, con x_1 se indica que $x = 1$ y x_0 se indica $x = 0$. Cuando un odds ratio se usa para asegurar que tener x_1 es dos veces más probable que x_0 de desarrollar alguna enfermedad, o que pacientes teniendo x son 40 % más posibles de morir dentro de 48 horas de los síntomas iniciales que los pacientes sin x, esto es lenguaje de cociente de riesgo (risk ratio) no lenguaje de cociente de odds (odds ratio). Estrictamente hablando, un modelo de regresión logística que ha sido parametrizado para estimar cocientes de odds debe ser interpretado como asegurando, por ejemplo, que los chances del resultado son dos veces mayores para x_1 comparado a x_0, no dos veces más probable. Pero hay veces que una interpretación de un odds ratio como riesgo es justificada.
En este capítulo se considera como usar los estimados de los coeficientes de regresión logística para hacer inferencias útiles en investigaciones epidemiológicas. Se consideran 3 tipos diferentes de variable x: binaria, cuantitativa, categórica. Cada caso se ilustra con una situación.

4.3. Factores de riesgos binomiales

Ejemplo 4.3.1. *Se consideran de nuevo los datos del Proyecto Pooling, cuadro 3.2, realizado en 1978 en cuatro ciudades de USA para estudiar los factores de riesgo de enfermedad cardiaca coronaria entre los hombres. En el cuadro 4.1 (recuerdo de estos datos) se presentan el status de fumador a la entrada y los casos de eventos coronarios durante el seguimiento, aproximadamente, 10 años para 1905 hombres con una edad entre 50-54 al inicio del estudio.*

	¿Evento coronario durante el seguimiento?		
¿Fumador al entrar?	Si	No	total
Si	166	1176	1342
No	50	513	563
Total	216	1689	1905

Cuadro 4.1: Fumadores y eventos coronarios en el Proyecto Pooling

Usando PROC GENMOD, uno de los procedimientos en SAS que puede usarse para ajustar modelos de regresión logística (el programa de computador está dado en el apéndice A). Se obtienen los resultados que se presentan en el cuadro 4.2. Por tanto el modelo

$$\widehat{logit} = -2{,}3283 + 0{,}3704x \tag{4.5}$$

ha sido ajustado. Para interpretar esto es necesario saber que los códigos usados para status de fumador, la variable x, fueron

$$x = \begin{cases} 1 & \textit{para fumadores} \\ 0 & \textit{para no fumadores} \end{cases}$$

Por [4.2] los odds ratio para un evento coronario, comparando fumadores y no fumadores es

$$\exp\{0{,}3704(1-0)\} = \exp(0{,}3704) = 1{,}448$$

Así los odds ratio para exposición comparado a no-exposición es particularmente fácil cuando la exposición se codifica con 1 y

Parámetro	Estimado	Error estándar
Intercepto(β_1)	$-2{,}3283(b_1)$	0,1482
Fumador(β_2)	$0{,}3704(b_2)$	0,1698

Cuadro 4.2: Extracto derivado de la salida SAS para los datos del proyecto Poolings

no-exposición como 0 *en el modelo de regresión logística. Todo lo que se debe hacer es calcular* $\exp(b_2)$.
Del cuadro 4.2, también tenemos que el error estándar de los ***log odds ratio*** $(\log \psi)$ *es* 0,1698. *Por [4.4] un intervalo de confianza del* 95 % *aproximado para* ψ *puede calcularse como:*

$$\exp\{0{,}3704 \pm 1{,}96 \times 0{,}1698\} = (1{,}038, 2{,}020)$$

Se puede también usar [4.5] si se desea, para hallar los odds de los datos del proyecto considerado. Los odds de un evento coronario resultan directamente de [4.5], como el logit es el log odds.
Para fumadores $(x = 1)$, *estos son*

$$\exp(-2{,}3283 + 0{,}3704 \times 1) = \exp(-1{,}9579) = 0{,}1412$$

Dando odds de 0,1412.
Para no fumadores $(x = 0)$, *los odds son*

$$\exp(-2{,}3283 + 0) = 0{,}0975$$

Dando odds de 0.0975.
Se puede estimar el riesgo, π, *de un evento coronario para fumadores de* [4,1] *y [4.5]. Esto es,*

$$\widehat{\pi} = \{1 + \exp(2{,}3283 - 0{,}3704 \times 1)\}^{-1} = 0{,}1237$$

De hecho, es algo más fácil de observar que

$$\widehat{\pi} = \left\{1 + \exp(-\widehat{logit})\right\}^{-1} \tag{4.6}$$

Y así, para no fumadores se tiene que el riesgo estimado es

$$\widehat{\pi} = \{1 + \exp(2{,}3283)\}^{-1} = 0{,}0889$$

El riesgo relativo estimado para fumadores comparado a no fumadores se halla entonces fácilmente como el cociente: $0{,}1237/0{,}0889 = 1{,}39$
Errores estándar, y límites de confianza, para odds generalmente no son sencillos de obtener. Esto debido a que pueden involucrar más de un parámetro y como los parámetros no son independientes, las covarianzas entre parámetros deben ser tenidas en cuenta. Por ejemplo, la varianza del $logit$ *cuando* $x = 1$ *(fumadores en el ejemplo considerado) es*

$$V(b_1) + V(b_2) + 2C(b_1, b_2)$$

	b_1	b_2
b_1	0,02195	-0,02195
b_2		0,02882

Cuadro 4.3: Matriz de Varianza-Covarianza para los datos del proyecto Pooling

Donde las Vs denotan varianzas y C covarianzas. SAS PROC GENMOD y otros software de computador pueden aplicarse para producir la matriz de varianza-covarianza (algunas veces simplemente llamando matriz de covarianza) de parámetros estimados. Esta es una matriz cuadrada con las filas y columnas marcadas por el parámetro.
Los elementos de la diagonal son las covarianzas de cada parámetro con si mismo, esto es, las varianzas.

Los elementos que no están el la diagonal son las covarianzas: como los elementos que están por encima y por debajo de la diagonal son iguales, solo uno de estas partes necesita ser reportada. La matriz de varianza y covarianza de los datos del proyecto Pooling se dan el cuadro 4.3. Por esto,

$$\widehat{V}(\widehat{\text{logit}}_{\text{fumadores}}) = 0{,}02195 + 0{,}02882 + 2 \times (-0{,}02195) = 0{,}00687$$

Y así

$$\widehat{se}(\widehat{\text{logit}}_{\text{fumadores}})) = \sqrt{0{,}00687} = 0{,}0829$$

Y el intervalo de confianza de 95 % para los odds de un evento coronario entre fumadores es

$$\exp\{-1{,}9579 \pm 1{,}96 \times 0{,}0829\} = (0{,}120, 0{,}1666)$$

los errores estándares de riesgos y riesgos relativos son difíciles de obtener porque ellos son funciones no lineales de los parámetros del modelo b_1 y b_2. En general, la varianza de una función no lineal no es la misma función de las varianzas. Por ejemplo, aun si se conoce $V(\text{riesgo}_{\text{fumadores}})$ y $V(\text{riesgo}_{\text{(no-fumadores)}})$ no se puede hallar fácilmente la varianza λ, del riesgo relativo para fumadores comparado a no-fumadores, como

$$V(\lambda) \neq V(\text{riesgo}_{\text{fumadores}}) \diagup V(\text{riesgo}_{\text{(no-fumadores)}})$$

Como consecuencia, solo errores estándares aproximados de riesgo y riesgo relativo puede derivarse. Aun estos son fastidiosos de calcular, luego es mejor usar odds ratio (y odds, si se requieren) como medidas de resultados si se usa regresión logística, cualquiera sea la forma de la variable exploratoria.

4.4. Factores de riesgo cuantitativos

Ejemplo 4.4.1. *Los datos del cuadro 4.4 muestran el número y porcentaje de muertes en un promedio de 7,7 años de seguimiento de un Estudio de Salud Cardiovascular Escocés (SHHS) realizado con 5754 hombres entre 40 − 59 años de edad.*

Edad (años)	Número de muertos	Total	Porcentaje de muertes
40	1	251	0.4
41	12	317	3.8
42	13	309	4.2
43	6	285	2.1
44	10	236	4.2
45	8	254	3.1
46	10	277	3.6
47	12	278	4.3
48	10	285	3.5
49	14	276	5.1
50	15	274	5.5
51	14	296	4.7
52	19	305	6.2
53	36	341	10.6
54	26	305	8.5
55	21	276	7.6
56	28	325	8.6
57	41	302	13.6
58	38	260	14.6
59	49	302	16.2

Cuadro 4.4: Número y porcentaje de muerte por edad en el estudio SHHS

Los datos se ilustran en la Figura 4.1. Es evidente que existe una tendencia a que el riesgo de muerte aumenta

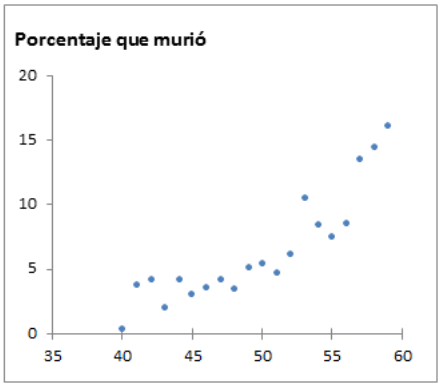

Figura 4.1: Porcentaje que murió contra edad al inicio del estudio en el estudio SHHS

con la edad, como seria de esperar, pero algunos grupos de edad individuales tienen más (o menos) de las muertes que esperaría de acuerdo con el patrón general.
Presumiblemente, esto es debido a la variación aleatoria. Modelos de regresión deben ser útiles tanto para resumir la relación que se muestra en el cuadro 4.4 y para suavizar la variación aleatoria.
Los parámetros estimados al usar SAS PROCGENMOD con estos datos se presentan en el cuadro 4.5. Así el modelo ajustado es:

$$\widehat{\text{logit}} = -8{,}4056 + 0{,}1126x \tag{4.7}$$

Donde el logit es el log odds de muerte y x es la edad del hombre. La interpretación de $b_2 = 0{,}1126$ es simplemente

Parámetro	Estimado	Error estándar
Intercepto (β_1)	$-8{,}4056(b_1)$	0,5507
Edad (β_2)	$0{,}1126(b_2)$	0,0104

Cuadro 4.5: Resultados producidos por SAS para los datos del estudio (SHHS)

la pendiente de la línea de regresión, completamente análoga a la situación en regresión lineal simple. Muestra que para cada aumento en la edad de 1 año, se estima que el logit aumenta en 0,1126. La figura 4.2 ilustra los valores logit ajustados y observados y los porcentajes. La regresión logística ha suavizado las irregularidades (particularmente a los 40 de edad) en los datos observados. Esto puede ser fácil de ver de los logits, donde los valores ajustados necesariamente siguen una línea recta. Sin embargo, la interpretación es más inmediata para los porcentajes. Los logit observados se calculan utilizando la definición de odds (3.8) los porcentajes ajustados (riesgos multiplicados por 100) salen de (4.1) al utilizar las estimaciones de los parámetros.
Para entender mejor las implicaciones prácticas del modelo ajustado se debe transformar los logits ajustados. Por ejemplo si planteamos la pregunta ¿Cuál es el odds ratio de los hombres mayores de 59 años con relación a los de edad de 40? Por [4.2] y [4.7], este es:

$$\widehat{\psi} = \exp\{0{,}1126(59-40)\} = 8{,}49$$

Hombres de 59 años tiene aproximadamente 8,5 veces más chance de morir en los próximos 7,7 años que los de edad de 40. Para calcular a 95 % de confianza para ψ se usa [4.4], sustituyendo el valor estimado de $se(b_2) = 0{,}0104$

$$\exp\{0{,}1126(59-40) \pm 1{,}96(59-40)0{,}0104\} = (5{,}77; 12{,}51)$$

Así estamos 95 % confiados de que el intervalo 5,77 a 12,51 contiene el odds ratio verdadero.

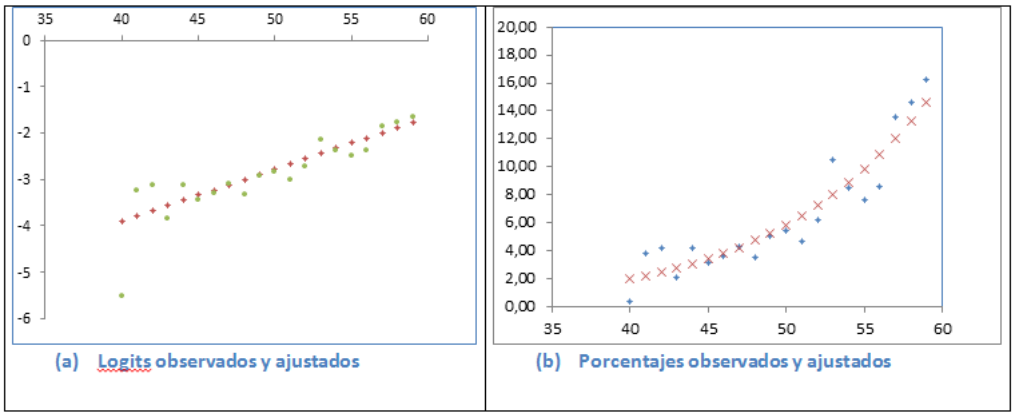

Figura 4.2: (a) logits observados y ajustados (b) porcentajes de un modelo de regresión logística para los datos del cuadro 4.4. Los valores observados son indicados por puntos, y los ajustados por cruces.

4.5. Factores de riesgo categórico

Cuando la variable x es categórica con l niveles, el modelo de regresión logística es ahora

$$\widehat{\text{logit}} = b_0 + b_1^{(1)}x^{(1)} + b_1^{(2)}x^{(2)} + \cdots + \cdots b_1^{(l)}x^{(l)} \tag{4.8}$$

donde cada variable $x^{(i)}$ es definida como

$$x^i = \begin{cases} 1 & \text{si x toma su nivel i} \\ 0 & \text{en otro caso} \end{cases}$$

Esto es las variables $\left\{x^{(i)}\right\}$ son variables mudas, las cuales juntas representan la variable x. En efecto, como x tendrá $l-1$ grados de libertad estas $x^{(i)}$ variables no pueden ser independientes, ya que hay l de ellas. Así, ellas deben ser ajustadas, sujetas a una restricción lineal arbitraria. Por [4.8], el logit estimado para el nivel i es

$$\widehat{\text{logit}}^{(i)} = b_0 + b_i^{(1)}$$

Sujeto a la restricción lineal escogida.
Entonces los odds ratio para el nivel i comparado al nivel j, el cual se denotará por $\psi^{(ij)}$, puede hallarse casi del mismo modo como [4.2] fue obtenido de [4.1]. Esto es

$$\widehat{\text{logit}^{(i)}} - \widehat{\text{logit}^{(j)}} = b_1^{(i)} - b_1^{(j)}$$

Luego

$$\widehat{\psi}^{((ij)} = \exp\left(b_1^{(i)} - b_i^{j}\right) \tag{4.9}$$

Además

$$\text{se}\left(\widehat{\text{logit}^{(i)}} - \widehat{\text{logit}^{(j)}}\right) = \sqrt{V(b_1^{(i)}) + V(b_1^{(j)}) - 2C(b_1^{(i)}, b_1^{(j)})} \tag{4.10}$$

Desafortunadamente, esto requiere el conocimiento de la covarianza $C\left(b_1^{(i)}, b_1^{(j)}\right)$. En el caso especial cuando $b_1^{(j)} = 0$, [4.9] y [4.10] se reducen a formas mucho más simple

$$\widehat{\psi}^{(ij)} = \exp\left(b_1^{(i)}\right) \tag{4.11}$$

$$\text{se}\left(\widehat{\text{logit}^{(i)}} - \widehat{\text{logit}^{(j)}}\right) = \text{se}\left(b_1^{(i)}\right) \tag{4.12}$$

Y si se usa [4.10] o [4.12] el intervalo de confianza de 95 % para $\psi^{(ij)}$ es

$$\exp\left\{(b_1^{(i)} - b_1^{(j)}) \pm 1{,}96\hat{s}e(\widehat{\text{logit}^{(i)}} - \widehat{\text{logit}^{(j)}})\right\} \tag{4.13}$$

Ejemplo 4.5.1. *Los datos del cuadro 4.6 muestran la presencia por Helicobacter pylori y la clase social ocupacional entre los hombres, en el tercer estudio MONICA en el norte de Glasgow.*
Los datos del cuadro 4.6 son para una variable exploratoria (clase social) con seis niveles categóricos. En tales casos hay varias maneras para definir los odds ratios, para resumir la relación entre factor riesgo y status de enfermedad, pueden definirse por ejemplo, cualquiera de los seis niveles de clase social como nivel básico, y todas las otras clases serían comparadas con esta referencia. Esto da lugar a cinco odds ratios de prevalencia. Cualquier otra escogencia de base daría lugar a cinco odds ratios diferentes. Cuando las variables categóricas se ajustan a paquetes de computador, la escogencia natural del nivel base varia según

Clase social ocupacional (Rango)	Número		Proporción con H. pylori
	Con H. pylori	Total	
I Profesional, No-manual (1)	10	38	0.26
II Intermedio, No-manual (2)	40	86	0.46
IIIn Experto, No-manual (3)	36	57	0.63
IIIm Experto, manual (4)	226	300	0.75
IV Parcialmente experto, manual (5)	83	108	0.77
V No calificado, manual (6)	60	73	0.82

Cuadro 4.6: Presencia de *Helicobacter pylori* y clase social ocupacional, entre los hombres, en el tercer estudio MONICA en el norte de Glasgow

el paquete escogido. Los datos de cuadro 4.6 fueron ajustados en SAS PROC GENMOD entrando el rango de clase social del cuadro 4.6 como la variable RANK, declarada como variable categórica (variable «CLASS» en notación SAS). Los parámetros estimados se presentan en el cuadro 4.7. El modelo ajustado de acuerdo a [4.8] es

$$\widehat{logit} = 1{,}5294 - 2{,}5590x^{(1)} - 1{,}6692x^{(2)} - 0{,}9904x^{(3)} - 0{,}4129x^{(4)} - 0{,}3294x^{(5)} + 0x^{(6)}$$

donde

$$x^{(i)} = \begin{cases} 1 \text{ si el rango de la clase social es } 1 \\ 0 \text{ si no} \end{cases}$$

Son las variables ficticias correspondientes y el paquete informático a proporcionado valores para los $\left\{b_1^{(i)}\right\}$. *SAS fija el parámetro para el ultimo nivel de cualquier variable categórica en cero. Esto implica que la restricción lineal impuesta por SAS sobre [4.8] es* $b_1^{(l)} = 0$. *En consecuencia utilizando SAS los odds ratio se calculan con relación con el nivel más alto. Esto es la clase social con rango* $= 6$ *como base. Así, por ejemplo, el odds ratio contraste clase social II (rango 2) a la clase social V (rango 6) es por [4.11]*

$$\widehat{\psi}^{(26)} = \exp(-1{,}6692) = 0{,}188$$

Con intervalo de confianza de 95 %, *de [4.12] y [4.13]*

$$\exp\left\{-1{,}6692 \pm 1{,}96 \times 0{,}3746\right\} = (0{,}90, 0{,}393)$$

De forma análoga se pueden obtener otros odds ratio tomando como base la clase social V (rango 6). Supóngase que se prefiere

Parámetros	Estimados	Error estándar
Intercepto	$1{,}5294b_0$	0,3059
RANK 1	$-2{,}5590b_1^{(1)}$	0,4789
RANK 2	$-1{,}6692b_1^{(2)}$	0,3746
RANK 3	$-0{,}9904b_1^{(3)}$	0,4111
RANK 4	$-04129b_1^{(4)}$	0,3340
RANK 5	$-0{,}3294b_1^{(5)}$	0,3816
RANK 6	$0{,}0000b_1^{(6)}$	0,0000

Cuadro 4.7: Resultados producidos pos SAS para el ejemplo 4.2

calcular odds ratio usando la primera clase social como base. Por ejemplo considere comparar la clase social II (rango 2) a la clase social I (rango 1) por [4.9]

$$\widehat{\psi^{(21)}} = \exp\left(b_1^{(2)} - b_1^{(1)}\right) = \exp\left(-1{,}6692 - (-2{,}5590)\right) = e^{0{,}8898} = 2{,}43$$

Un modo alternativo de obtener este mismo resultado de la tabla de estimadores de parámetros es notar que

$$\begin{aligned} \widehat{\psi^{(21)}} &= \widehat{\psi}^{(26)}\widehat{\psi}^{(61)} = \widehat{\psi}^{(26)} / \widehat{\psi}^{(16)} \\ &= \exp(-1{,}6692) / \exp(-2{,}5590) \\ &= \exp\left(-1{,}6692 - (-2{,}5590)\right) \end{aligned}$$

El mismo resultado de antes.
Para obtener el intervalo de confianza de 95 % *para* $\widehat{\psi^{(21)}}$ *se debe obtener primero la matriz de varianza-covarianza. La matriz presentada en el cuadro 4.8 producida por SAS PROC GENMOD. Note que* $b_1^{(6)} = 0$, *es fijado por SAS y así no tiene varianza. Así por [4.10]*

$$\hat{se}\left(\widehat{logit}^{(2)} - \widehat{logit}^{(1)}\right) = \sqrt{0{,}14033 + 0{,}22930 - 2 \times 0{,}09359} = 0{,}4271$$

Por tanto, el intervalo de confianza de 95 % *para* $\widehat{\psi}^{(21)}$ *es, por [4.13]*

$$\exp\left\{-1{,}6692 - (-2{,}5590) \pm 1{,}96 \times 0{,}4271\right\} = \exp\left\{0{,}8898 \pm 0{,}8371\right\} = (1{,}05, 5{,}62)$$

Hay una gran ventaja en tener un parámetro «pendiente» de cero para el nivel de base escogido. Como consecuencia, es posible forzar el paquete de computador a hacer esto, y así se evitaría que los cálculos sean más extensos. Por ejemplo si el nivel base del ejemplo anterior se hubiese escogido $b_1^{(1)} = 0$, el estimado y el error estándar estimado de $\psi^{(21)}$ saldrían directamente de estos paquetes. Algunos paquetes (incluyendo GENSTAT y GLIM) en contraste con el SAS, fijan el primer nivel a tener un parámetro «pendiente» de cero $b_1^{(1)} = 0$.
La forma más fácil de forzar el paquete escogido para usar una base que no sea la que el paquete automáticamente escoja cuando los niveles se entran en su orden numérico natural, es presentar, al paquete los niveles de la variable x ordenados apropiadamente. Algunos paquetes, tales como SPSS permiten al usuario un grado de escogencia al colocar el nivel base.

	b_0	$b_1^{(1)}$	$b_1^{(2)}$	$b_1^{(3)}$	$b_1^{(4)}$	$b_1^{(5)}$
b_0	0.09359	-0.09359	-0.09359	-0.09359	-0.09359	-0.09359
$b_1^{(1)}$		0.22930	0.09359	0.09359	0.09359	0.09359
$b_1^{(2)}$			0.14033	0.09359	0.09359	0.09359
$b_1^{(3)}$				0.16899	0.09359	0.09359
$b_1^{(4)}$					0.11153	0.09359
$b_1^{(5)}$						0.14564

Cuadro 4.8: Matriz de varianza-covarianza producida por SAS para el ejemplo 2.2

4.6. Datos genéricos

Hasta ahora se ha asumido que los datos son dados al paquete de computador utilizado en forma agrupada (como en el cuadro 4.9). Sin embargo los datos reales son usualmente no agrupados, simplemente recordando el nivel de factor de riesgo y el status de la enfermedad para cada individuo separadamente. Esto es el formato de datos genérico (o caso por caso, o binario) como se muestra en el cuadro 4.10. En la práctica, los datos genéricos son

Valor del factor de riesgo	Número con enfermedad	Número total	Proporción con enfermedad
x_1	e_1	n_1	r_1
x_2	e_2	n_2	r_2
.	.	.	.
.	.	.	.
x_l	e_l	n_l	r_l

Cuadro 4.9: Datos agrupados de valores de los factores de riesgo y resultado de la enfermedad

Valor del factor de riesgo	¿Enfermedad?
x_1	si
x_2	no
.	.
.	.
.	.
x_n	no

Cuadro 4.10: Los datos brutos sobre los valores de los factores de riesgo y resultado de la enfermedad

más fáciles de tratar que los datos agrupados. Es simplemente más fácil describir la metodología usando el formato agrupado. Con datos genéricos se codifica la variable status de enfermedad como 1 si la enfermedad está presente y

0 si no. Se puede ajustar modelos de regresión logística justo como antes excepto que en la notación del cuadro 4.10. , $n_i = 1$ para cada i(cada individuo)

$$e_i = \begin{cases} 1 \text{ si el individuo } i \text{ tiene la enfermedad} \\ 0 \text{ si no} \end{cases}$$

Y así, $r_i = e_i/n_i = e_i$.
Por ejemplo el cuadro 4.4 presenta una versión resumida, agrupada, de los datos SHHS sobre edad y muerte. La base original para el estudio es, conceptualmente, una matriz rectangular de individuos contra variables. Cuando se seleccionan las variables de status edad y muerte se obtiene una matriz con 5754 filas y dos columnas (EDAD y MUERTE, codificadas como 0 para no, 1 para si). Se necesita definir una tercera columna que consta de 5754 entradas del numero 1. Estas tres columnas toman el lugar de las primeras tres columnas en la tabla 4.4, aunque edad tendrá ahora muchos valores duplicados y no estarán ordenados. Estas tres columnas de datos genéricos pueden ser entradas a un paquete de computador y ser ajustado un modelo de regresión logística.

4.7. Modelos de regresión logística múltiple

Al igual que con los modelos de regresión estándar, modelos logísticos se pueden especificar con varias variables explicativas, en lugar de sólo una. Teniendo en cuenta k variables explicativas, el modelo de regresión logística múltiple es por analogía con [4.1]

$$\log\left(\frac{\hat{r}}{1-\hat{r}}\right) = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

donde $\hat{r} = \hat{\pi}$ es el riesgo estimado de enfermedad. El lado derecho de la ecuación del modelo es ahora un predictor lineal múltiple.
Si, por ejemplo, x_i es una variable categórica con l niveles, entonces se reemplaza b_ix_i con

$$b_i^{(1)}x_i^{(1)} + b_i^{(2)}x_i^{(2)} + \cdots + b_i^{(l)}x_i^{(l)}$$

Donde

$$x_i^{(j)} = \begin{cases} 1 \text{ si } x_i \text{ toma su nivel } j \\ 0 \text{ si no} \end{cases}$$

Sujeta a una restricción lineal arbitraria (tal como $b_i^{(j)} = 0$). Todo esto, aparte de las diferencias identificadas en la sección 4.1 es exactamente como para modelos generales. La regresión logística es un ejemplo de un modelo lineal generalizado.
Los modelos logísticos múltiples se tratan casi del mismo modo que los modelos de una variable, pero (como con modelos lineales generales de variable múltiple), permiten un rango más amplio de inferencias.

Ejemplo 4.7.1. *El cuadro 4.11 presenta datos sobre enfermedad coronaria (CHD) presión sanguínea y colesterol de una base cohorte del SHHS. Se está considerando esto como un estudio de cohorte fijo con un periodo de seguimiento de 7.7 años. Los datos se muestran para 4093 hombres sin evidencia de CHD al inicio del estudio, para los cuales la presión arterial sistólica y el colesterol total en suero fueron medidos. La presión sanguínea y los grupos de colesterol escogidos en el cuadro 4.11. es la quinta parte de los datos de todo el conjunto de hombres, es decir, antes de que las personas con cardiopatía coronaria CHD al inicio del estudio fueron eliminados.*
Estos datos se leen utilizando el paquete SAS, denotando la variable presión sanguínea sistólica (SBP) como SBP5TH y la variable colesterol total como CHOL5TH. Al analizar los datos en PROC GENMOD, fijando el nivel más bajo de SBP5TH y CHOL5TH respectivamente a tener un odds ratio de cero (odds de unidad). Los parámetros estimados producidos por SAS se dan en el cuadro 4.12, utilizando estos valores el modelo ajustado sería:

$$\begin{aligned} \widehat{logit} = &-4{,}5995 + 0x_1^{(1)} + 0{,}6092x_1^{(2)} + 0{,}8697x_1^{(3)} + 1{,}0297x_1^{(4)} + 1{,}3425x_1^{(5)} + 0x_2^{(1)} \\ &+ 0{,}2089x_2^{(2)} + 0{,}8229x_2^{(3)} + 1{,}0066x_2^{(4)} \\ &+ 1{,}2957x_2^{(5)} \end{aligned} \tag{4.14}$$

	Total colesterol sérico (mmol/l)				
Presión arterial (SBP)	≤ 5.41	5.42-6.01	6.02-6.56	6.57-7.31	> 7.31
≤ 11 8	1/190	0/183	4/178	8/157	4/132
119-127	2/203	2/175	6/167	10/166	11/137
128-136	5/173	9/176	9/181	8/167	11/164
137-148	5/139	3/156	10/154	13/174	16/174
> 148	5/123	8/123	12/144	13/197	23/180

Cuadro 4.11: Relación de eventos de CHD con el número total de presión arterial sistólica (SBP) y quintas partes de colesterol para hombres en los SHHS que estaban libres de enfermedad coronaria (CHD) al inicio del estudio

donde $x_1^{(i)}$ representa el nivel i de SBP y $x_2^{(i)}$ representa el nivel i del colesterol, para $i = 1, \cdots, 5$ el orden de rango en cada caso. Como los estimados mas altos corresponden a los niveles más altos en cada caso ($b_1^{(5)}$ y $b_2^{(5)}$), se ve inmediatamente que es muy peligroso estar en los niveles quintos de SBP y colesterol.
Se puede estimar los log odds de enfermedades del corazón CHD para un hombre en el nivel más alto de SBP ($> 148\, mmHg$) y más alto nivel de colesterol ($> 7{,}31\, mmol/l$) para obtener, de [4.14]

$$\widehat{logit} = \{-4{,}5995 + 1{,}3425 + 1{,}2957\} = -1{,}9613$$

los odds son entonces $e^{-1{,}961} = 0{,}1407$.
Se puede estimar el riesgo para un hombre en estos dos niveles extremos usando [4.6]

$$\widehat{\pi} = \widehat{r} = \left\{1 + \exp(-\widehat{logit})\right\}^{-1} = \{1 + \exp(1{,}9613)\}^{-1} = 0{,}1233$$

Así que la probabilidad de un evento coronario en el periodo de seguimiento de 7,7 años es de 0,1233 para hombres de mediana edad con niveles altos de SBP y colesterol.
Se puede también estimar el log odds ratio para un hombre en la combinación de altos niveles $(5,5)$ relativo a la combinación de niveles más bajos $(1,1)$ ($SBP \leq 118 mmHg; colesterol \leq 5{,}41 mmol/l$) por [4.14] como:

$$\begin{aligned}\log \hat{\psi} &= \widehat{logit}^{(5,5)} - \widehat{logit}^{(1,1)} \\ &= (-4{,}5995 + 1{,}3425 + 1{,}2957 - (-4{,}5995)) \\ &= 1{,}3425 + 1{,}2957 = 2{,}6382\end{aligned}$$

Luego el odds ratio, $\widehat{\psi}$, es $e^{2{,}6382} = 14{,}0$. Nótese que el término constante $(-4{,}5995)$ simplemente se cancela cuando un odds ratio se calcula. El riesgo relativo para el mismo contraste es

$$\hat{r}^{(5,5)} / \hat{r}^{(1,1)}$$

Que resulta ser $0{,}1233/0{,}009957 = 12{,}4$ Nótese que esto es similar a los odds ratio, como seria de esperar ya que CDH es razonablemente inusual, aun en el grupo de alto riesgo.
Miremos como se comportan los dos factores de riesgo, cuando se varían solos para compararlo con el efecto combinado ya visto. Por el cuadro 4.12, el odds ratio comparando el nivel 5 de SBP al nivel 1, manteniendo el colesterol fijo es $e^{1{,}3425} = 3{,}8$. Es perfectamente correcto ignorar todos los estimados para colesterol al derivar esto. Para verlo, considere el odds ratio para el 5° comparando al primer quinto para SBP manteniendo el colesterol fijo en el tercer quinto. Esto es lo mismo que hallar el odds ratio para la combinación de (SBP, colesterol) como $(5,3)$ relativa a $(1,3)$. De [4.14]

$$\begin{aligned}\widehat{logit}^{(5,3)} &= -4{,}5995 + 1{,}3425 + 0{,}8229 \\ \widehat{logit}^{(1,3)} &= -4{,}5995 + 0 + 0{,}8229\end{aligned}$$

Parámetros		Estimados
Intercepto		−4,5995
SBP5TH	1	0,0000
SBP5TH	2	0,6092
SBP5TH	3	0,8697
SBP5TH	4	1,0297
SBP5TH	5	1,3425
CHOL5TH	1	0,0000
CHOL5TH	2	0,2089
CHOL5TH	3	0,8229
CHOL5TH	4	1,0066
CHOL5TH	5	1,2957

Cuadro 4.12: Estimados para el ejemplo 4.7.1 producidos por SAS

Por lo tanto los log odds son $\widehat{logit}^{(5,3)} - \widehat{logit}^{(1,3)} = 1{,}3425$ *y* $\widehat{\psi} = e^{1,3425}$*, como se afirmó. El término constante y el parámetro pendiente para el nivel 3 de colesterol se han cancelado. Similarmente, para comparar entre niveles extremos de colesterol, manteniendo SBP fijo, el odds ratio es* $e^{1,2957} = 3{,}7$*. Note que* $3{,}8 \times 3{,}7 = 14{,}1$*, lo cual es muy similar al odds ratio combinado de* 14,0 *hallado anteriormente, sugiriendo que hay poca interacción entre estos dos factores de riesgo.*
Los intervalos de confianza para comparaciones donde solo niveles de una variable varían se sigue exactamente como en la sección 4.4. Cuando los niveles de dos (o más) variables varían se necesitara la matriz varianza-covarianza. Por ejemplo, el error estándar de el log odds ratio comparando $(5,5)$ *a* $(1,1)$ *en el ejemplo 4.7.1 es*

$$\begin{aligned} se\left\{\widehat{logit}^{(5,3)} - \widehat{logit}^{(1,3)}\right\} &= se\left\{b_0 + b_1^{(5)} + b_2^{(5)} - \left(b_0 + b_1^{(1)} + b_2^{(1)}\right)\right\} \\ &= se\left\{b_1^{(5)} + b_2^{(5)}\right\} \\ &= \sqrt{V\left(b_1^{(5)}\right) + V\left(b_2^{(5)}\right) + 2C\left(b_1^{(5)}, b_2^{(5)}\right)} \end{aligned}$$

Usando esto, el intervalo del 95 % *de confianza es, como ya es usual es*

$$estimado \pm 1{,}96\hat{s}e$$

Capítulo 5

Diseño de estudio para explorar comportamiento de fumar y esquizofrenia

5.1. Conceptos epidemiológicos y estadísticos

5.1.1. Adición

En este trabajo se usa el termino «adicción» para expresar la naturaleza compulsiva de tomar una droga adictiva (Volkov y Fowler 2000). La adicción a la droga es un grupo de trastornos crónicos recurrentes caracterizados por la compulsión a tomar la droga, perdida de control en limitar la toma y emergencia de estados emocionales negativos cuando no hay acceso a la droga (Koob y Volkov 2010).
El sistema de diagnóstico Psiquiátrico oficial en USA se llama Manual estadístico y diagnostico de desórdenes mentales y es una variante de la tercera edición publicada en 1980, llamada DSM-III (Asociación Psiquiátrica Americana 1980). La edición actual de la DSM continúa usando la terminología «Trastorno por consumo de sustancia» distinguiendo entre abuso (uso patológico) y dependencia (uso patológico severo con síntomas de adicción). Esta terminología puede cambiar en la próxima edición para expresar el sentido de «adicción» (O'Brien et al. 2006) usando el termino «uso de sustancias y desórdenes adictivos».
Actualmente no hay una definición oficialmente aceptada de adicción, pero según Goodman, (1990) adicción designa un proceso donde o por el cual un comportamiento que puede funcionar para producir placer y proveer escape a descontrol interno se emplea en un patrón caracterizado por (1) fallo recurrente de controlar el comportamiento (sin poder) y (2) continuación del comportamiento a pesar de consecuencias negativas significativas (inmanejabilidad).
La tercera y las ediciones DSM subsiguientes usan la palabra «sustancias» para significar «nicotina», «alcohol», y otras sustancias de abuso usualmente incluidas bajo el nombre de «drogas». El termino «drogas» usualmente se refiere a «drogas ilegales», o drogas prescritas usadas inapropiadamente. Así, en este trabajo a no ser que específicamente sea necesario (por ej. «dependencia de nicotina», la cual es específicamente definida) usa los términos «adicción a alcohol», «adicción a nicotina» y «adicción a drogas» (refiriéndose a drogas distintas a nicotina y alcohol). Cuando se usa la palabra adicciones en plural quiere decir adicción a cualquiera de estas sustancias en general.
Para investigadores interesados en genética, es importante saber que los estimados de heredabilidad de adicción son $40-60\,\%$ (Volkow y Li 2005) o $50-60\,\%$ (Bierut 2011). Bierut (2011) recientemente describió la vulnerabilidad genética a adicción como probablemente reflejando la combinación de cientos de miles de genes de efectos modestos.
Un asunto complejo no resuelto completamente en la literatura es que los factores genéticos que influyen en el uso de las drogas pueden ser diferentes de aquellos que influyen en la adicción. Estudios paralelos sugieren que la iniciación y patrones tempranos del uso de las drogas son fuertemente influenciados por factores ambientales, sociales y familiares mientras que los patrones posteriores de uso y niveles de adicción son altamente influenciados por factores genéticos (Kendler et al.2008). Aun si se acepta la hipótesis de que factores genéticos pueden ser mucho más influyentes en las adicciones a droga que en el uso de droga, una explicación del rol importante de influencias genéticas sobre adicción sigue siendo compleja.
Una hipótesis biológica propone que los factores genéticos se expresan una vez expuestos a suficientes cantidades de la droga sobre un periodo suficientemente largo. La hipótesis socialmente mediada es basada en el hallazgo de que nuestros genes juegan un rol creciente en formar nuestro propio ambiente social, esto propone que los genes pueden influir en la selección de ambientes que activan o encaucen el uso de droga (Kendler et al.2008)

5.1.2. Adición a Nicotina

Estudios epidemiológicos en los 1980s y los 1990s en los EE.UU y otros países occidentales establecieron que la mayoría de los fumadores en la población general fuman al menos un cigarrillo al día; esto es, son fumadores diarios. Pocas personas en la población general de los países occidentales se consideran fumadores ocasionales. La nicotina es el componente aditivo más importante del mundo del tabaco. Así, fumar diario es usualmente considerado como signo de adicción a nicotina. En la población general, iniciación a fumar diario raramente ocurre en la década de los 20s (Nelson et al.1998; De León y et al.2002; Dierker et al.2008). Fumadores ocasionales no fuman diariamente, parecen no tener mayores problemas (no síndrome de abstinencia) cuando no fuman y así, no son adictos (Shiffman 1989). «Fumar Diario» es una medida simple y razonable de adicción a la nicotina en los estudios epidemiológicos en países occidentales.

En países subdesarrollados, pocos individuos se pueden dar el lujo de gastar en fumar diariamente los fumadores no-diarios es necesario incluir en estudios sobre el fumar. El status socioeconómico necesita ser cuidadosamente también considerado (Campo-Arias et al.2006). En la población en U.S. restricciones a fumar cada vez mayores han sido implementadas desde el 2000. Esto ha llevado a un aumento en la proporción de los fumadores no-diarios que ahora son un tercio de los fumadores de USA en la población general (Shiffman 2009). Por tanto Shiffman (2009) ha propuesto que fumar diario en cantidad puede ocurrir solo cuando fumar no es restringido, por restricciones económicas, sociales o legales.

El DSM que se ha vuelto ampliamente usado internacionalmente desde su tercera edición (DSM-III) no lista «abuso de nicotina» como un diagnostico posible; solo «dependencia de nicotina» es listada. Así el abuso y conceptos de dependencia de este sistema clasificatorio nunca ha servido particularmente bien para la adicción a nicotina. Las razones pueden ser que: 1) Cuando se compara con otras drogas, la nicotina es una sustancia particularmente aditiva (Hennigfeld et al.1991; López-Quintero et al.2011); y 2) Es una sustancia legal cuyo uso descontrolado está muy raramente asociado con complicaciones legales. En un gran estudio epidemiológico en USA, López-Quintero y sus colaboradores (2011) estiman que el 68 % de los usuarios de nicotina se vuelven dependientes (versus 23 % para alcohol, 21 % para cocaína y 9 % para los usuarios de cannabis). Los conceptos de DSM de dependencia de nicotina sin embargo han sido usados en unos pocos estudios epidemiológicos de la población general pero no en pacientes con esquizofrenia.

Los clínicos e investigadores que trabajan en el cese de fumar frecuentemente usan una escala para definir dependencia de nicotina. El test Fagerstrom para dependencia de nicotina (FTND) es una medida altamente usada de dependencia de nicotina, que ha sido probada para predecir razonablemente el éxito en dejar de fumar (Heatherton et al.1991). Esta escala tiene seis items. Sin embargo, dos de los seis items, los cuales tienen un puntaje entre 0 y 3 pueden reflejar mejor la dependencia de nicotina: Ítem 1 (El tiempo entre despertar y fumar el primer cigarrillo del día), y ítem 4(el número de cigarrillos fumados por día). La suma de los puntajes de estos dos ítem es llamado el índice de fumar alto (HSI) (Heatherton et al1998). Un FTND o HSI altos parecen ser una buena indicación de alta dependencia de nicotina (De León et al.2003b; Díaz et al. 2005). Restricciones de fumar institucionales pueden artificialmente disminuir estos índices (Steinberg et al. 2005).

5.1.3. Esquizofrenia y adicción a nicotina

Otra afirmación infundada, repetida por muchos artículos es que el $80-90$ % de pacientes de esquizofrenia fuman (Chapman et al.2009). Unos pocos estudios realizados en los 1980s y los 1990s en países occidentales, dieron estas prevalencias muy altas. Sin embargo estos estudios no son representativos de pacientes de esquizofrenia alrededor del mundo y reflejan un punto en el tiempo cuando el fumar era aun usado como terapia de refuerzo en los hospitales.

Además, aunque estamos convencidos de que la asociación entre la esquizofrenia y fumar es muy consistente alrededor del mundo y una explicación biológica probablemente la justificará, se debe reconocer que, como con cualquier otra droga aditiva, la adicción a nicotina en la esquizofrenia es obviamente influida por la disponibilidad de la droga. Así, la prevalencia de fumar tabaco en una muestra particular de pacientes de esquizofrenia va a ser influida por la disponibilidad de tabaco en el lugar particular y tiempo en el cual se recolecta la muestra. La revisión hecha de 42 muestras en 20 países da una prevalencia combinada presente de fumar del 62 %. Más importante que esta prevalencia promedio es el tamaño de la asociación entre esquizofrenia y fumar cuando se compara pacientes de esquizofrenia con la población general. Este efecto se mide con un meta-análisis global OR (y su CI), 5.3 (4.9-5.7) (de León, Díaz y Quintana 2005). Este promedio estimado de OR no es una medida «verdadera» y estable de la asociación. El promedio mundial OR debe ser recalculado como nuevos estudios reflejan cambios en prevalencias de fumar en la población general de cada país, y cambios en el acceso a tabaco de la población que va a desarrollar

esquizofrenia (de León et al. 2007a).
La epidemia de tabaco alrededor del mundo usualmente pasa por 5 etapas (temprano, subiendo, pico, declinante, últimos estados). Diferentes países están usualmente en diferentes etapas, y las mujeres tienden a estar más temprano que los hombres. Como mucha gente de la población de US han suspendido el fumar, los US están actualmente en periodo declinante. Los países de bajos salarios están actualmente en las primeras etapas (Anderson 2006). La población con esquizofrenia tiende a estar un poco aislada de las tendencias de reducción de fumar (de León et al. 2002b) y después de muchos años de estudios de personas con esquizofrenia en varios países occidentales muchos estudios revelan tasas muy bajas de cesación de fumar (menor 20 %) en pacientes de esquizofrenia en condiciones naturales, (ósea sin tratamiento) las cuales son probablemente típicas en la mayoría de contextos psiquiátricos.

5.1.4. Asociación entre SMIs y adicciones

El concepto de enfermedad mental severa (SMI) que además abarca otras enfermedades, usualmente incluye esquizofrenia y otros varios desórdenes mentales severos (desórdenes bipolar, y desórdenes depresivo grave) y pueden abarcar casi el 3 % de la población en US en un periodo de un año (Consejo de Salud Mental Nacional 1993). En US, la población con esquizofrenia y varios otros desórdenes mentales son tratados frecuentemente en el sistema de salud publica.
En el apéndice A cuadro 1, describe cuatro modelos generales que han sido usados para explicar la asociación entre adicciones y SMIs (Mueser et al. 1998). Obviamente múltiples modelos pueden explicar una asociación especifica (Gregg et al. 2007). Basado en un tratamiento epidemiológico (de León 1996), propone que un modelo de vulnerabilidad genética compartida puede explicar la alta prevalencia de fumadores entre la población con esquizofrenia. Como otros datos verificaron que esta asociación es consistente alrededor del mundo, y que el aumento en la iniciación a fumar comienza antes de la esquizofrenia, se ha llegado a creer firmemente que la vulnerabilidad compartida refleja un componente genético (de León y Díaz 2005). Desde una perspectiva diferente que considera datos neurobiológicos, Freedman y colaboradores (1997) defienden el punto de vista de que esta asociación puede ser explicada por una asociación entre esquizofrenia con variaciones genéticas en el gene receptor de nicotina $\alpha 7$. Esta explicación genética no ha sido empíricamente confirmada pero más recientemente, Freedman y sus colaboradores hallaron anormalidades o anomalías en la explicación de estos receptores en no fumadores esquizofrénicos quienes al ser comparados con controles, las anomalías no estaban presentes en fumadores con esquizofrenia (Mexal et al. 2010)

5.2. Diseño de estudio para explorar comportamiento de fumar y esquizofrenia

Esta sección se enfoca en dos asuntos principales que son cruciales para diseñar estudios de la asociación entre esquizofrenia y fumar: la definición de comportamiento de fumar y la selección de controles.

5.2.1. Definición de comportamiento de fumadores

De un modo simplificado, la asociación entre esquizofrenia y el fumar tabaco puede significar: 1) la frecuencia de fumadores es mayor en pacientes de esquizofrenia que en otras personas («un porcentaje mayor de ellos fuman») y/o 2) los fumadores esquizofrénicos fuman más cigarrillos al día que otros fumadores («ellos fuman más cigarrillos»).
Se pueden dar más significados precisos por medio de hipótesis contrastables que comparan pacientes con esquizofrenia con controles como sigue: 1) «Esquizofrenia es asociada con un aumento en fumar», y 2) « Esquizofrenia es asociada con fumar en aumento entre los fumadores».
El fumar diario actual refleja dos procesos de adicción: 1) siempre ser adicto, y 2) persistencia de adicción (falta de cesar de fumar en los fumadores diarios). Así la asociación entre esquizofrenia y fumar diario y fumar diario actual puede ser particularmente útil para aquellos interesados en completar estudios genéticos de esquizofrenia y adicción a nicotina. Hechos más interesantes son la asociación estadística de esquizofrenia con frecuencias aumentadas de fumar, y disminución de cesar de fumar entre fumadores diarios.

5.2.2. Selección de controles

Sobre el estudio de sujetos de control, dos tipos principales de control pueden seleccionarse: controles no-psiquiátricos representando la población general y paciente con otras SMIs. Los pacientes con otras SMIs son un grupo de comparación mejor que la población general en estudios de pacientes de esquizofrenia (ver tabla A.2 apéndice A para argumentos). Esta bien establecido que los desórdenes psiquiátricos son asociados con frecuencias altas de fumar (Hughes et al. 1986; Covey et al. 1994; Lasser et al. 2000). Esto es particularmente valido para personas con SMIs distintos de esquizofrénicos, tales como desorden depresivo grave o desorden bipolar, quienes exhiben mayores frecuencias de fumar que los controles no-psiquiátricos en todos los estudios epidemiológicos en US (Covey et al. 1994; Lasser et al. 2000; McClave et al. 2010). No es sorprendente que solo unos pocos estudios usan pacientes con otros SMIs como estos estudios usualmente tienen mucho menos poder estadístico que los estudios usando controles no-psiquiátricos. En una encuesta realizada en el 2007 en US la diferencia en prevalencias ajustadas a la edad de fumadores actuales entre esquizofrenia y desorden bipolar fue 12,7 %$(= 59{,}1 - 46{,}4)$; en contraste una diferencia de 40,8 %$(= 59{,}1 - 18{,}3)$ entre pacientes de esquizofrenia y controles no psiquiátricos fue observada (McClave et al. 2010).

5.3. Esquizofrenia asociada con más fumadores

Esta sección se enfoca en el diseño de estudios de sondeo de la asociación entre esquizofrenia con más fumadores. Seis tipos de estudios son posibles incluyendo aquellos enfocados en: 1) un aumento en fumar presente con respecto a la población general; 2) un aumento en fumar corriente con respecto a otros SMIs ; 3) un aumento en el fumar con respecto a la población general; 4) un aumento en fumar con respecto a otros SMIs; 5) una disminución en dejar de fumar en fumadores con respecto a la población general; o 6) una disminución en dejar de fumar en fumadores con respecto a otros SMIs.

¿Esquizofrenia es asociada con un aumento en fumar presente al comparar con la población general?

Como se dijo anteriormente, las prevalencias de fumar en la población general varían en tiempo y espacio. También a través de las muestras de pacientes de esquizofrenia, pero tal vez menos. Así, es más importante concentrarse en los tamaños del efecto, cuantificando la diferencia en prevalencia entre pacientes de esquizofrenia y controles y su estabilidad.
El meta análisis realizado mostró una muy fuerte asociación entre esquizofrenia y fumar $(OR = 5{,}3)$ ver tabla A.3 apéndice A. ¿Fue esta asociación consistente entre los países? Si 40 de los 42 estudios revisados mostraron una asociación significativa en la misma dirección (OR>1, ósea frecuencia de fumar más grande en esquizofrenia) solo dos estudios dieron $OR < 1$: un estudio Colombiano mostrando también una baja prevalencia de fumar en la población general (Suarez et al. 1996), y un estudio Japonés con altas prevalencias de fumar (Mori et al 2003). Estos dos estudios reflejan un «efecto de piso» y un «efecto de techo» respectivamente. La asociación entre esquizofrenia y fumar no puede hallarse si nadie fuma en la población general $(0\,\%\text{piso})$ o si todos fuman $(100\,\%\text{techo})$. Revisando estudios a nivel mundial (de León, díaz, et al. 2005) los efectos piso y techo pueden ocurrir en alrededor de 20 % y 60 % de fumar actual en la población general, respectivamente (de León et al. 2007a). Así, cuando solo 20 % de la población son fumadores, es difícil hallar una asociación a no ser que las variables confusas como genero y status socio-económicos sean cuidadosamente controladas y muestras grandes usadas. Una muestra de esquizofrenia Colombiana que no mostró diferencias significativas al compararse con las tasas publicadas para la población general (Campo et al. 2004) se volvieron significativamente diferentes al compararse con controles $(OR = 3{,}1, CI, 1{,}4 - 6{,}8)$ (Campo-Arias et al. 2006). En el tiempo del estudio Japonés (Mori et al. 2002) pocas mujeres fumaban $(13\,\%)$ y muchos hombres fumaban $(60\,\%)$ en la población general Japonesa. Así, el estudio de esquizofrenia Japonés que no mostró una asociación entre fumar y esquizofrenia (Mori et al. 2003) estaba contaminado por usar tasas publicadas de la población general y no controladas por factores de confusión y diferencia de genero. De hecho, en ese estudio Japonés, las mujeres esquizofrénicas fumaban significativamente más que la población general (de León, Díaz 2005) y la asociación entre esquizofrenia y fumar en fumadores hombres estuvo limitada por el efecto techo de la población general Japonesa masculina. Un estudio Japonés más reciente mostró que después de controlar factores de confusión era también posible demostrar una asociación entre esquizofrenia y fumar en Japón (Shinozaki et al. 2010).
Cambios en la epidemia de fumar han sido muy influidos por factores económicos y de genero (Anderson 2006). Para presentar una versión simplificada de la cronología de la epidemia de tabaco debe notarse que en los países occidentales, las diferencias de género eran obvias en la primera mitad del siglo 20 como pocas mujeres fumaban,

pero las diferencias han disminuido desde entonces. En países no occidentales, el fumar de mujeres continúa siendo raro pero es creciente, y la mayor parte del negocio del tabaco parece estar moviéndose hacia hombres Asiáticos (Anderson 2006).
Si se cree que las diferencias de genero en el consumo de tabaco son importantes, es necesario reconocer que ORs usando la población general como control son medidas pobres de asociación si no se ajustan para genero. Sin embargo ORs estratificadas por genero puede ser la mejor manera de describir la asociación entre esquizofrenia y fumar alrededor del mundo.
En el meta-análisis realizado, el ORs masculino global para fumador fue 7.2 ver (Tabla A.3 apéndice A). Más importante aun, de los 32 estudios que proporcionaron ORs masculinos, tres dieron $ORs < 1$; sin embargo estos tres estudios fueron llevados a cabo en países no occidentales, probablemente con efecto techo. Muestras de acuerdo a status educativos y socioeconómicos pueden ser necesarias para detectar la asociación en pacientes con efectos techo.
En el meta- análisis el OR femenino global fue 3.3 (ver Tabla A.3 apéndice A.). Cinco de los 25 estudios que proporcionaron ORs femeninos dieron $ORs < 1$. Estos estudios se hicieron en países occidentales con efectos piso; de nuevo, estudios pueden ser necesarios para detectar la asociación (de León y Díaz 2005).

¿La esquizofrenia puede ser asociada con un aumento en fumar al compararse con otros SMI?

Cuando se compara la esquizofrenia con otros SMIs, el meta-análisis realizado produjo un OR global de 1.9 (de León Díaz 2005). El OR masculino fue 2.3 y OR femenino fue 1.8 pero solo 3 de ocho estudios tuvieron significativos ORs mayor que 1 (ver Tabla A.3 apéndice A). Desafortunadamente, pocos estudios comparando esquizofrenia con otros SMIs han controlado por otras adiciones. En países occidentales el alcohol y adiciones a la droga están fuertemente asociados con el fumar. En US y otros países occidentales, otras adiciones pueden ser asociadas con SMIs en general y con esquizofrenia en particular. Así, estudios solidos metodológicamente tratan de establecer la relación entre esquizofrenia y fumar, necesitan controlar para efectos confusos de genero y desórdenes de uso de sustancias. Pocos estudios publicados han controlado estos efectos confusores. (de León y otros 2002; 2005).

¿La esquizofrenia asociada con aumento de fumar si se compara con la población general?

El OR global del meta-análisis fue 3.1 (de León y Díaz 2005). El OR de hombres fue 7.3 y el de mujer fue 2.8 (ver Tabla A.3 apéndice A) los estudios revisado no controlaron por confusos tales como el nivel de educación. En un estudio más reciente en U.S usando fumar diario como definición y controles no-psiquiátricos, hubo una asociación significativa de fumar con esquizofrenia después de controlar por confusos $(OR = 5{,}2 \quad CI \quad 3{,}6 - 78)$ (de León y et al. 2007a). Este OR grande puede ser parcialmente explicado por factores asociados con esquizofrenia que no son bien controlados usando controles no psiquiátricos (por ejemplo la presencia de otras adicciones). Usando la población general como control no permite la posibilidad de ajustar para otras adicciones .Como otros niveles de adicción en la población general son bajos, y así, difícil de representar en muestras de población general.
Mediante el uso de análisis de supervivencia enfocados en la edad de inicio del fumar diario, y controlando por factores confundidores, se ha específicamente explorado la sincronización en la asociación de esquizofrenia con una iniciación mayor de fumadores diarios en tres estudios (dos en USA y uno en España). Los tres estudios (de León y otros 2002a; Gurpegui et al. 2005; Díaz et al. 2008), replicaron que: 1. Cuando se comparan con controles no-psiquiátricos, la esquizofrenia es asociada con un aumento significativo en la iniciación de fumar diario después de controlar para educación y genero; 2. Fumadores no-psiquiátricos mostraron un patrón de volverse adictos en la adolescencia con poca gente adicta en los 20 tardíos, en contraste, los pacientes de esquizofrenia continúan en riesgo de volverse adictos después de la edad de 20. Enfocado solo en adultos que empezaron a fumar diario 5 años antes de la aparición de la esquizofrenia, dos de los estudios (Gurpegui et al. 2005; Díaz et al. 2008) demostraron que las diferencias entre pacientes de esquizofrenia y controles no psiquiátricos no se explicaron por los efectos prodrómicos de la enfermedad esquizofrénica. Un estudio en pacientes Chinos con esquizofrenia replico la asociación de la esquizofrenia con aumento en la iniciación de fumar diario enfocándose en individuos que comenzaron a fumar cinco años antes de la aparición de la enfermedad.(Zhang et al. 2010).

¿Es la esquizofrenia asociada con aumento de fumar diario al compararse con otros SMIs?

El OR global del meta-análisis fue 2.0 y significativo para el total de muestras masculinas, pero no alcanzo a ser significativo en muestras femeninas (ver Tabla A.3 apéndice A.) Las tasas medias de fumar de por vida en los pacientes con esquizofrenia fue 69 % para las muestras totales, 83 % para hombres y 65 % para mujeres (de León y Díaz

2005). Estos estudios no controlaron por algunos confundidores tales como nivel de educación y desórdenes de uso de substancias.
En un estudio mas reciente de fumadores, hubo una asociación significativa de fumar diariamente con esquizofrenia después de controlar por confundidores $OR = 1{,}9(CI1{,}2 - 2{,}8)$ (De León y otros 2007a)

¿Es la esquizofrenia asociada con disminución de fumar en fumadores cuando se compara con la población general?

El OR global del meta-análisis de falta de cesar de fumar para las muestras totales de fumadores fue 5.3 los ORs masculinos y femeninos fueron 10.0 y 2.2 respectivamente (Cuadro A.3). Estos estudios no controlaron por otros confundidores. En un estudio mas reciente en U.S que definió dejar de fumar como no fumar por un año después de haber sido fumador, que usó controles no-psiquiátricos, hubo una asociación significativa entre esquizofrenia y una falta de cesar de fumar después de controlar por confundidores $(OR = 5{,}6; CI3{,}4 - 9{,}1)$ (de León y otros 2007a).
Los predictores más fuertes de no dejar de fumar son, probablemente, fumar en exceso y alta dependencia a la nicotina. Como la esquizofrenia esta altamente asociada con fumar y posiblemente con alta dependencia de nicotina (Ver la próxima sección), la explicación mas razonable para la asociación entre esquizofrenia y falta de cesar de fumar son los altos niveles de dependencia de nicotina en pacientes de esquizofrenia. Antes de considerar factores genéticos como una posible explicación para la disminución en cesar de fumar en fumadores con esquizofrenia, se necesita considerar la asociación entre esquizofrenia y alta dependencia de nicotina en fumadores.
Estudios Naturalísticos indican que la cesación de fumar en fumadores esquizofrénicos es baja $(< 10\,\text{o}\,20\,\%)$ cuando no se provee ningún tratamiento especial. Debido a los altos niveles de dependencia, se necesita programas intensivos para ayudar a los pacientes de esquizofrenia a parar de fumar (Ziedonis y George 1997). Según la literatura de ensayos clínicos, las únicas intervenciones que han resultado un poco efectivas en reducir el fumar en esquizofrenia son bupropion y refuerzo contingente (tsoi; 2010).

¿Es esquizofrenia asociada con disminución en dejar de fumar cuando se compara con otros SMIs?

El numero limitado de estudios y sus tamaños muéstrales limitados hacen difícil sacar una conclusión confiable sobre la asociación entre dejar de fumar y esquizofrenia al comparar pacientes de esquizofrenia con otros SMIs (Tabla 5). En el mas reciente y minucioso estudio en U.S el cual uso controles con otros SMIs y definen cesar de fumar en un fumador diario como no haber fumado por un año, no hubo una asociación significativa entre falta de dejar de fumar y esquizofrenia después de controlar por confundidores $(OR = 1{,}2; CI, 0{,}68 - 2{,}1)$ (de León et al 2007a).

5.4. ¿Es esquizofrenia asociada con fumar más en los fumadores?

Un articulo viejo (Lohr and Lynn 1992) describe experiencias medicas que los fumadores esquizofrénicos parecen fumar más cigarrillos y ser más adictos que otros fumadores. Desde entonces esta afirmación ha sido continuamente repetida en la literatura. Esta sección se centra en el diseño de los estudios que prueban la hipótesis de que la esquizofrenia se asocia con más consumo de tabaco en los fumadores. Cuatro tipos de estudios son posibles enfocándose en: 1. Aumento en fumar en fumadores esquizofrénicos con respecto a fumadores en la población general; 2. Aumento en el fumar en fumadores de esquizofrenia con respecto a los fumadores con otros SMIs; 3. Altos niveles de dependencia de nicotina en fumadores esquizofrénicos comparados a fumadores en la población general; 4. Altos niveles de dependencia de nicotina en fumadores esquizofrénicos comparados a fumadores con otras SMIs. Una hipótesis algo relacionada que se revisa en la quinta subsección es si el fumar en esquizofrenia es asociado con más bajo pronóstico en esquizofrenia o no.

¿Es esquizofrenia asociada con aumento bastante en fumar entre fumadores cuando se compara con la población general?

Los estudios publicados han usado diferentes definiciones de fumar, haciendo imposible proveer o dar un OR promedio. Sin embargo, tienden a sugerir que los fumadores con esquizofrenia consistentemente fuman más que los fumadores en la población general. En el meta-análisis, el OR masculino varia de 2,0 a 7,4 y el OR femenino de 2,0 a 8,8 (de León y Díaz 2005).Un estudio controlado por género, educación, raza y edad $(OR = 2{,}9, CI(1{,}5, 5{,}6)$; Díaz et al.2008) indicó que el fumar bastante fue significativamente más frecuente en fumadores con esquizofrenia $(42\,\%)$ que en fumadores sin desórdenes psiquiátricos $(14\,\%)$.

Solo unos pocos estudios han utilizado métodos biológicos para establecer la posibilidad de que los fumadores con esquizofrenia puedan tener niveles de metabolismo de nicotina más alto que otros fumadores sin desórdenes psiquiátricos (Olincy et al.1997; Strand y Nyback 2005; Weinberger et al 2007; Williams et al 2005, 2010). Puede ser difícil establecer definitivamente este hecho a no ser de que se completen estudios más amplios de muestras representativas de fumadores con esquizofrenia y fumadores sin desórdenes psiquiátricos y confundidores sean controlados. No es sorprendente que en un estudio de niveles de orina no halla diferencias después de controlar por el numero de cigarrillos fumados como ellos estaban controlando para intensidad de fumar (Bozikas et al. 2005).

¿Es esquizofrenia asociada con aumento de fumar en fumadores al comparar con otros SMIs?

El uso de diferentes definiciones de fumar en exceso (fumadores pesados) ha hecho imposible proveer un OR promedio al comparar fumadores en exceso con esquizofrenia versus fumadores con otros SMIs. Las diferencias tienden a ser pequeñas y no significativas, con aproximadamente $2/3$ ORs mayores que 1 y $1/3$ de ellos menores que uno. En un estudio más reciente en U.S la prevalencia de fumadores pesados dentro de los fumadores fue ligeramente más alta en esquizofrenia que en otros desórdenes (42 vs 33 %), pero la diferencia no fue significativa después de controlar por confundidores $(OR = 1{,}6; CI, (0{,}94, 2{,}6)$; Díaz et al. 2008) No hay estudios usando medidas biológicas comparando fumadores con esquizofrenia versus fumadores con otros SMIs.

¿Es esquizofrenia asociada con niveles más altos de dependencia de nicotina dentro de los fumadores al compararse con la población general?

Estudios controlando por confundidores y usando medidas de dependencia de nicotina basados sobre el FTMD (Gurpegui et al.2005; Díaz et al.2008) indican que los pacientes de esquizofrenia que fuman son probablemente más dependientes que los fumadores sin desorden psiquiátrico. Mas interesante son tres estudios menores (Tidey 2005, Lo et al. 2011; Williams et al.2011) reportando que cuando los paciente de esquizofrenia que fuman son controlados con otros fumadores altamente dependientes los primeros desean más la nicotina; esto es algo que ha sido sospechado por años (Mckee et al.2009; Weinberger et al.2007).

¿Es esquizofrenia asociada con niveles más altos de dependencia de nicotina dentro de fumadores al comparar con otros SMIs?

Actualmente se tiene buena información de varios países mostrando que los SMIs en general son consistentemente asociados con niveles más altos de dependencia de nicotina que los fumadores dentro de la población general (de león et al.2002c, 2003; Díaz Et al 2009), pero no hay datos definitivos mostrando que los fumadores con esquizofrenia fuman más y son más dependientes que los fumadores con otros SMIs. Las pequeñas diferencias observadas y la inconsistencia de los resultados indican que se necesitan muestras muy grandes para resolver esta pregunta.

En los países occidentales el fumar bastante y la alta dependencia de nicotina son también frecuentes en alcohólicos y otros usuarios de drogas. En un estudio de un programa de tratamiento para dejar de fumar, alcoholismo y esquizofrenia fueron asociados con pobres resultados (Gershon Grand et al. 2007). La única comparación publicada que se ha encontrado de dependencia de nicotina en fumadores con esquizofrenia versus fumadores tratados para alcohol y otras adicciones a drogas es un estudio canadiense en una unidad de pacientes (Solty et al. 2009). El estudio incluyo 84 pacientes psicóticos (Solo 73 %, $62/84$, con esquizofrenia o desorden Sico afectivo) y 31 con diagnósticos primario de adicción. La prevalencia (Sin corrección para genero) de fumar fueron 55 % y 77 %, y de alta dependencia (FTND mayor o igual que 6) fueron 47 % y 65 % respectivamente (Solty et al.2009).

¿El fumar es asociado con el peor pronóstico en pacientes de esquizofrenia?

Otro modo de explorar la relación entre esquizofrenia y fumar alto es comparar el resultado psiquiátrico de pacientes de esquizofrenia que son grandes fumadores con los pacientes que no lo son, controlando por otros factores que puedan influir el pronostico. Muchos de los estudios que dan altos porcentajes de fumar en esquizofrenia, $80 - 90$ % fueron completados en hospitales en un largo plazo. Esto sugiere que los pacientes institucionalizados quienes usualmente tienen los peores pronósticos fueron más propensos a fumar (de león y Díaz 2005). Los pocos estudios limitados que han explorado sistemáticamente la asociación entre fumar y esquizofrenia sus resultados indican que los fumadores (Comparados a no fumadores) o fumadores pesados (Comparados a fumadores suaves y no fumadores) tienden a tener pronóstico longitudinales más pobres (Aguilar et al.2005; Salokangas et al. 2006; Kobayashi et al 2010; Wang et al 2010; Segarra et al 2011). Estos resultados van directamente contra la idea de automedicación. Si el fumar es útil, los fumadores pesados, en exceso, debieran tener mejores pronósticos que los fumadores suaves y los no fu-

madores. Muchos más estudios controlando por variables confundidoras se necesitan para establecer una asociación entre fumar o fumar pesado en fumadores y peores pronósticos. Si este hecho es establecido con nuevos estudios puede tener sentido considerar la hipótesis de allostasis. Kolb y Le Moal (1997) sugieren que el organismo trata de contrarrestar los efectos de una droga dada a través de un círculo vicioso en el cual el punto donde se logra el placer cambia continuamente en respuesta a la administración de la droga. Ellos afirman que la drogadicción resulta de la desregulación de mecanismos de recompensa y la subsiguiente allostasis (perturbaciones crónicas de la homeostasis de recompensa cerebral en la cual la estabilidad puede alcanzarse solo por cambios). Siguiendo el concepto de alistases fumar mucho en esquizofrenia puede ser una situación de no ganar asociada con un disturbio mayor de los sistemas compensatorios del cerebro los cuales son incapaces de ajustarse ellos mismos a pesar de los intentos de los paciente a fumar mucho mas para ajustarlos(De León et al 2005).

Capítulo 6

Aplicaciones

6.1. Iniciación a fumar y esquizofrenia: un estudio de replicación en una muestra Española

Resumen:
En un estudio anterior en US, la vulnerabilidad a la esquizofrenia se asocia con un mayor riesgo de iniciarse a fumar diariamente después de los 20 años de edad. Un análisis de supervivencia de edad de aparición de fumar diario compara 290 controles con 250 pacientes esquizofrenicos DSM-IV consecutivos de facilidades hospitalarias en un área urbana en España. Después de controlar genero y educación, las curvas hazard acumuladas para edad de iniciación a fumar y pacientes de esquizofrenia fueron significativamente diferentes. Después de la edad de los 20, las tasas de iniciar a fumar fueron mayores en todos los pacientes de esquizofrenia (y en 107 pacientes quienes comenzaron a fumar diariamente al menos cinco años antes de la aparición de la enfermedad).

1. **Introducción**.
   Cierta evidencia sugiere la hipótesis de que la esquizofrenia puede estar asociada con una mayor vulnerabilidad a comenzar a fumar (De león, 1996). Por ejemplo, la esquizofrenia se asocia a tasas de fumar relativamente altas, y muchos pacientes con esquizofrenia comienzan a fumar antes de la aparición de la enfermedad. Seis estudios conducidos en cuatro países diferentes hallaron que la prevalencia a fumar en pacientes de esquizofrenia eran significativamente mayores que en aquellos de la población general (cociente de odds promedio $OR = 3{,}8$, intervalo de confianza de $(95\,\%, CI = 2{,}8 - 4{,}8)$ (Martínez-Ortega, 2004).
   Además, al combinar cuatro estudios en cuatro países diferentes, prevalencia a fumar en los pacientes de esquizofrenia resulto significativamente mayor que la de otros pacientes con enfermedad mentalmente severa $(OR = 1{,}6 CI = 1{,}2 - 2{,}1)$ (Martínez-Ortega, 2004).
   Un estudio en US produjo mayores evidencias a la anterior hipótesis. Comparó la edad de aparición o iniciación a fumar diario (AODS) en pacientes de esquizofrenia versus la gente en la población general y versus pacientes con otras enfermedades mentales (De León, 2002) y sugirió que los individuos con esquizofrenia o vulnerabilidad a la esquizofrenia tiene mayor riesgo de iniciación a fumar diariamente después de los 20 años de edad. El estudio uso técnicas de análisis multivariado para controlar los efectos potencialmente confusos de genero y educación. El objeto del presente estudio fue replicar el anterior análisis de edad de aparición de fumar diario con muestras Españolas.

2. **Sujetos y métodos**.

2.1 Sujetos
El estudio se localizo en dos centros de salud mental y un programa de rehabilitación cubriendo la ciudad de Granada (sur de España). Todos los pacientes recibían tratamiento psiquiátrico gratis por el sistema de salud nacional (o seguro social), el cual se dividió en areas de captación incluyendo facilidades para pacientes internos y externos.
La muestra externa ha sido descrita previamente e incluyo los primeros 250 pacientes consecutivos diagnosticados con esquizofrenia DSM-IV (18 de 268 rechazados). La diagnosis fue hecha por un psiquiatra investigador

con la version clinica de una entrevista de diagnostico estructurada (First et al., 1994).
Todos los pacientes eran caucasianos. Le edad media fue 36,1 $(DS = 9{,}5$años$)$ y la edad en diagnosis fue de 21,9 $(DS = 6{,}0)$. La edad de un paciente diagnosis se define como la edad en la cual el paciente fue visto por primera vez con síntomas psiquiátricos que llevaron aun diagnostico final de esquizofrenia.
Fueron 195 hombres (78 %). Nueve por ciento (23/250) de los pacientes con educación universitaria. La mayoría de los pacientes (94 % 236/250) estaban tomando antisicóticos, muchos bajo antisicóticos típicos (71 %, 171/250). La dosis media de chlorpromazine equivalente (APA, 1997) en aquellos bajo antisicóticos eran 550mg/dia (DS=459).
Una muestra de controles sin enfermedades mentales psicóticas $(N = 290)$ incluyo pacientes medicos, pacientes de otros pacientes medicos, y empleados de otra clínica quienes fueron estudiados con el cuestionario de salud general (GHQ-28) (Lobo et al., 1986). Las enfermedades mentales psicóticas fueron excluidas después de preguntar sobre tratamiento psiquiátrico anterior o actual. Una persona con un desorden bipolar y tres con esquizofrenia fueron excluidas del grupo de control, pero personas con desorden mental no-psicótico fueron excluidas. La edad media de los controles fue 40.5 años y 113 controles fueron hombres (36 %). 49 % $(141/290)$ de los controles tenia educación universitaria. La prevalencia de fumar actual en los controles $(35\,\%$, tabla $1)$ era la misma, la que se hallo en un estudio reciente de la población general española (Pinilla y Gonzalez 2001). Todos los pacientes y controles dieron consentimiento escrito después de una descripción del estudio completo.

2.2 **Método**.
Como en el estudio anterior (de León et al., 2002), se preguntó a todos los pacientes establecer su historia de fumador. Se les preguntó sobre $1)$ fumar diario, definido como el fumar sobre base permanente durante algún periodo de vida; $2)$ fumar diario presente; $3)$ edad de aparición del fumador diario (AODS); y $4)$ dependencia de nicotina alta, definida como un puntaje ≥ 6 en el Test Fagerstrom para Dependencia de Nicotina (FTND) (Fagerstrom et al., 1996). La variable nivel de educación se divide en alta (educación universitaria) o baja (ver Tabla 1 apéndice B.)
Cocientes no ajustados (OR) que comparan pacientes esquizofrénicos con controles se calcularon por medio de cross-tabulation, y los ORs ajustados para género y educación por regresiones logísticas (Tabla 1 apéndice B). Como en el estudio de US (de León, 2002) la iniciación a fumar diariamente se exploró por medio de un análisis de supervivencia AODS. Para una edad particular, la tasa de iniciación a fumar se define como la tasa de hazard y se mide por la pendiente de la curva de hazard acumulada a esa edad (Figura 1 apéndice B). La tasa de hazard para una edad particular es la proporción por unidad de tiempo de la gente que nunca fumo antes de esa edad particular y que empezara a fumar a esa edad especifica (Klein y Moeschberger, 1997). Pruebas estratificadas Log-rank para diferencias entre las curvas acumuladas se llevo a cabo, controlando para genero y nivel de educación(Klein y Moeschberger, 1997; de León et al., 2002).
Para descartar la posibilidad de que estos resultados puedan ser explicados por cambios prodrómicos, iniciación de tratamiento psiquiátrico, imitación de pacientes o la institucionalización de los controles, la curva hazard acumulada de los controles fue comparada a la de aquellos pacientes que comenzaron a fumar antes del periodo prodrómico. La literatura usualmente considera que el periodo prodrómico dura hasta un año (Elkhaen et al., 2003). McGlashan (2003) asume que puede durar de uno a dos años. Bajo esta hipótesis, la curva hazard acumulada para los controles fue inicialmente comparada a la de aquellos pacientes que comenzaron a fumar al menos dos años antes de la aparición de la enfermedad para tomar en cuenta la posibilidad de que algunos pacientes puedan tener periodos prodrómicos mayores a dos años, esta comparación se repitió con solo pacientes de esquizofrenia que comenzaron a fumar al menos 3, 4, 5 años antes de la aparición de la enfermedad.

3. **Resultados**.
Como se esperaba, cuando se compara con los sujetos de control, los pacientes de esquizofrenia tuvieron prevalencias significativamente más alta de fumar diario $(OR = 2{,}1)$, fumar diario actual $(OR = 4{,}3)$ como también alta dependencia de nicotina $(OR = 8{,}1)$, y fumar en exceso $(OR = 3{,}9)$ entre los fumadores diarios actuales (Tabla 1 apéndice B). Después de ajustar por genero y educación, los OR para fumar diario presente en todos los sujetos, y para alta dependencia de nicotina en todos los fumadores fueron aun significativas (Tabla 1 apéndice B). Ambos OR ajustados y no ajustados fueron significativos entre hombres, mujeres e individuos con bajo o alto nivel educativo (Tabla 1 apéndice B).
La curva hazard acumulada para todos los pacientes de esquizofrenia fue significativamente diferente de los controles, aun después de controlar género y educación (Fig. 1. apéndice B). Después de 20 años de edad, los pacientes tuvieron tasas de iniciación significativamente más altas que los controles (ver Fig. 1 caption. apéndice B).

La curva hazard acumulada para pacientes que comenzaron a fumar al menos 5 años antes de la aparición de la enfermedad $(N = 107)$ continuó significativamente diferente de los controles aun después de controlar por genero y educación (Fig. 2 apéndice B). Como antes, después de 20 años de edad, tasas de iniciación para pacientes de esquizofrenia fueron significativamente mayores (Fig. 2 caption. apéndice B). Resultados similares se obtuvieron por al menos 4 años $(N = 118; \log-\text{rank estratificado} \chi^2 = 88{,}9, df = 1, p < 0{,}001)$ 3 años $(N = 129, \chi^2 = 84{,}9, df = 1, p < 0{,}001)$ y 2 años $(N = 136, \chi^2 = 85{,}2, df = 1, p < 0{,}001)$.

4. **Discusión**.
   La principal conclusión de este estudio es que si un no-fumador con esquizofrenia o vulnerable a la esquizofrenia es mayor de 20, la probabilidad de que el o ella se convierta en un fumador diario es más alta o mayor que la de un no-fumador de la población general.
   Esta conclusión sugiere que algunas personas vulnerables a la esquizofrenia, se convertirán en fumadores en sus 20, cuando otras personas raramente inician el fumar diario. La comparación AODS de controles y aquellos pacientes de esquizofrenia que comenzaron a fumar al menos 5 años antes de la aparición de la enfermedad sugiere que las diferencias observadas en tasas de iniciación entre controles y pacientes, probablemente no se deben a cambios prodromales o tratamiento psiquiátrico.
   Las conclusiones anteriores son consistentes con las del estudio US realizado antes (De Leon , 2002) y soportan la hipótesis de que una vulnerabilidad a la esquizofrenia puede estar asociada con incrementar una mayor vulnerabilidad a comenzar a fumar (De Leon 1996). En efecto, hay alguna información que sugiere que este puede ser el caso. Freedman et al. (1997) describen una anormalidad neuropsicológica genética en pacientes con esquizofrenia (y sus parientes), la cual es corregida temporalmente con un pico alto de nicotina. Esta anormalidad es asociada con una disfunción de un receptor de nicotina especifica del hipocampo $(\alpha 7)$. Mas recientemente, Leonard et al. (1998) hallaron que la presencia de un polymorphismo promotor $(\alpha 7)$ era más frecuente en pacientes con esquizofrenia que en controles y puede ser una marca de las anormalidades neuropsicológicas que aumentan el riesgo de esquizofrenia.
   Nuestros resultados difieren de los de un estudio de cohorte en hombres conscriptos suecos en edades de 18-20 que sugirió que el fumar puede tener un efecto protectivo contra el desarrollo de la esquizofrenia en pacientes con edad de aparición entre 20 y 25 años (Zammit et al., 2003). Este estudio fue limitado por la exclusión de mujeres y de todos los hombres que comenzaron la esquizofrenia antes de la conscripción, y por la falta de información sobre la iniciación a fumar en hombres mayores de 20, lo cual es crucial para establecer diferencias significativas entre controles hombres y pacientes de esquizofrenia hombres, según nuestros resultados.
   Nuestro análisis incluyendo pacientes con esquizofrenia que comenzaron a fumar 5 años antes de la aparición de la enfermedad, sugiere que la asociación entre fumar y esquizofrenia no puede ser explicada por la enfermedad o periodo prodronal. La literatura describe que algunos precursores de la esquizofrenia tales como deficit neurológicos menores pueden ya estar presentes en la infancia y algunos especulan que los cambios neuropatológicos tales como disturbios en migración de neuronas pueden ocurrir ya antes del nacimiento. Sin embargo, ni los precursores, ni los disturbios neuropatológicos prenatales específicos a la esquizofrenia y probablemente muchos sujetos que sufrieron estos disturbios nunca desarrollan esquizofrenia. Así, parece mejor considerar precursores y cambios neuropatológicos prenatales como factores de riesgo para la esquizofrenia. Obviamente, si se defiende el punto de vista extremo, no compartido, que la esquizofrenia incluye los precursores y los cambios neuropatológicos tempranos, ambos de estos ocurren mucho antes de la iniciación a fumar en pacientes con esquizofrenia y se tendría así que considerar la iniciación a fumar en pacientes con esquizofrenia como otro signo temprano de la enfermedad.

5. **Conclusión**.
   Esta replica Española de resultados US sugiere que la vulnerabilidad a la esquizofrenia, puede estar asociada con un riesgo mayor de volverse fumador diario. Estudios prospectivos de pacientes con primeros episodios psicóticos o sujetos con riesgo de desarrollar esquizofrenia puede requerirse para establecer mejor la interacción entre vulnerabilidad a esquizofrenia, AODS y edad de aparición de la esquizofrenia.

6.2. Menos pero mayores consumidores de cafeína en esquizofrenia: Un estudio de control-caso

Resumen:
Según la literatura, hay una asociación entre esquizofrenia y consumo de cafeína, pero no es claro si la esquizofrenia es asociada con o mayor prevalencia de toma de cafeína diaria o la cantidad consumida. En este estudio se compara nuestros pacientes de esquizofrenia previamente publicados ($n = 250$) con una muestra de control ($n = 290$), después de controlar por variables demográficas y consumo de tabaco y alcohol. Toma de cafeína actual fue menos frecuente en pacientes esquizofrénicos ($59\,\%, 147/250$) que en los controles ($70\,\%, 204/290$). En los análisis multivariados la toma de cafeína fue menos frecuente a menor edad y en pacientes de esquizofrenia, y más frecuente en fumadores y usuarios de alcohol. Entre los consumidores de cafeína, toma de cafeína bastante (mayor $o = 200mg/dia$) fue significativamente asociada con esquizofrenia(64 % en esquizofrenia vs 36 % en controles) como también mayor edad y fumar. Cantidad diaria de toma de cafeína y cigarrillo fumados se correlacionan significativamente en el grupo de esquizofrenia pero no en el de control. La correlación entre toma de cafeína y dependencia de nicotina fue baja y no significativa en ambos grupos. La asociación entre fumar y toma de cafeína alta puede ser parcialmente explicada por un efecto farmacokimetrico: Los compuestos de fumar tabaco inducen metabolismo de cafeína por el citocrome P450 1 A2. Aunque la esquizofrenia en si misma puede ser asociada con toma de cafeína alta en los usuarios de cafeína, parte de esta asociación se explica por la asociación entre esquizofrenia y fumar. La asociación entre toma de cafeína y alcohol parece ser mas compleja, uso de alcohol y cafeína fueron asociados significativamente, pero dentro de los usuarios de cafeína, el alcohol fue asociado con consumo de cafeína menos frecuente entre los fumadores. En estudios posteriores la medición de niveles de plasma de cafeína ayudara a definir mejor toma de cafeína alta y a controlar por efectos farmacokimetricos de fumar.

1. **Introducción**.
   La literatura entre cafeína y esquizofrenia es limitada, ya que incluye principalmente reporte de casos o estudios que no controlan para confusores(Schneier y Siris, 1987; Hugles et al.,1998). En su resumen anterior Schneir y Siris (1987) sugieren que la esquizofrenia puede estar asociada a una tasa mayor de consumo de cafeína en nuestro estudio reciente de 250 pacientes de esquizofrenia españoles después de controlar las variables confusoras como alcohol y tabaco, se hallo que la toma de cafeína estaba asociada con fumar y uso de alcohol, pero no con sintomatología esquizofrénica (incluyendo ansiedad y depresión) o la dosis de medicación antipsicótica. (Gurpegui et al, 2004) entre los consumidores de cafeína esquizofrenicos el fumar se asocia con la toma de cafeína alta. Una comparación cruzada con los datos limitados de dos reportes de cafeína en la población española en general sugirió que los pacientes de esquizofrenia parecen tener prevalencias de tomas de cafeína diarias similares, pero la proporción de pacientes con una cantidad alta (mayor $o = 200mg/dia$) de toma de cafeína ($94/250$) parece ser alrededor del 10 % mas alta que en la población general.
   En nuestro estudio previo no se tuvo en cuenta controles para comparar con pacientes de esquizofrenia después de controlar factores con fusores y así no se logro establecer significativamente si o no los pacientes esquizofrénicos tienen o poseen un patrón diferencial de consumo de cafeína. El objetivo de esta extensión fue comparar nuestros pacientes previamente publicados con una muestra de control después de controlar las variables demográficas y el uso de tabaco y alcohol en un diseño control-caso.

6.3. Sujetos y métodos

6.3.1. Sujetos y procedimiento

Los pacientes y controles proveen consentimiento informado escrito después de una descripción completa del estudio; los protocolos fueron aprobados por la junta institucional del hospital universitario San Cecilio (Granada, España). La muestra de los pacientes de esquizofrenia estable ha sido previamente descrita. Incluyo los primeros 250 pacientes consecutivos diagnosticados con esquizofrenia DSM-IV. Su edad (media desviación estándar) $36{,}1 \pm 9{,}5$ años. 195 hombres ($78\,\%$) y 10 % de ellos ($24/250$) con educación universitaria.
El grupo de control para este estudio ha sido ya descrito. Brevemente, los sujetos fueron reclutados en una clínica de pacientes de medicina de familia. Los 290 individuos eran de 18 años de edad o más, e incluyeron pacientes ($58\,\%$), parientes de pacientes ($34\,\%$) y empleados($8\,\%$) cuatro sujetos fueron excluidos debido a sufrir de un desorden psicótico (uno de desorden bipolar y tres de esquizofrenia). Una versión Española previamente validada del

cuestionario de salud general de 28 ítems, una herramienta frecuentemente usada en Europa, fue usada para detectar posibles casos con síntomas psiquiátricos. Usando un puntaje > 6 (sensibilidad $77-89\,\%$; especificidad $86-90\,\%$), 84 de los 290 controles $(29\,\%)$ fueron o resultaron posibles casos psiquiátricos, incluyendo 32 sujetos quienes reportaron haber sido diagnosticados con desórdenes psiquiátricos no-psicóticos; 35 estaban tomando la droga psicotrópica. Esta prevalencia de 29 % es similar a la de 28,6 % hallado en un estudio de población general en la provincia de Granada. La prevalencia de fumadores $(35\,\%)$ fue similar a la prevalencia de fumadores en la población general reportada en la literatura $(34{,}5\,\%)$. La edad media de los controles fue $40{,}5 \pm 15{,}1$ años 39 % hombres y 49 % con educación universitaria. En resumen, la muestra de control en este estudio parece ser una representación razonable de la población general de Granada.
Además de las características socio-demográficas de los participantes, su consumo de bebidas cafeinadas (café, té y colas), tabaco, alcohol y drogas ilegales fue considerado. La toma de cafeína semanal promedio se determinó estimando el contenido de cafeína en las bebidas y luego convirtiendo a mg/da. El contenido de cafeína standard utilizado para el café fue 100mg en una taza de café de 150cc, 45 mg en una tasa de te; y para las sodas cafeinadas 23 mg en un vaso de soda de 200 cc. Basados en tasas de consumo de cafeína para la población general, se definió una toma de cafeína alta como 200 mg o más por día. La evaluación de fumar ha sido ya descrita, incluyo el test Fagerstrom para dependencia de nicotina FTND. El uso corriente de alcohol y drogas ilegales se evaluó por un médico investigador en controles y un psiquiatra en los pacientes de esquizofrenia. En controles, las evaluaciones de alcohol y drogas se llevó acabo por entrevistas semi estructuradas; en los pacientes se llevó a cabo por medio de entrevistas semi-estructuradas, revisión de opiniones e información colateral de la familia.

6.3.2. Análisis estadístico

Los análisis se llevaron a cabo con el paquete estadístico para ciencias sociales (SPSS 12). Los cocientes ($OR's$) y sus intervalos de confianza de 95 % (CI) se calcularon por tabulación en cruz-a-dos para análisis univariado. Un análisis con regresión logística con variables independientes dicotómicas se efectuó. La variable dependiente dicotómica fue la presencia - ausencia de toma de cafeína corriente, y las variables independientes, fueron la enfermedad de esquizofrenia, el fumar diario actual, uso de alcohol, genero, edad mayor de 37 años, y nivel de educación (esta variable fue partida entre alta, educación universitaria, y/o baja). Un análisis de regresión logística similar se efectuó con los consumidores de cafeína una toma de cafeína alta ($\geqq 200mg/da$) como la variable dependiente. Todos los modelos logísticos se ajustan bien según el test de bondad de ajuste Hosmer-Lemeshow (H. L. 2000). Las correlaciones de Spearman se usaron para analizar entre la cantidad de toma de cafeína y otras variables. Los test de Kruskal-Wallis se usaron para comparar las cantidades de cafeína en esquizofrenia y los sujetos de control que consumieron cafeína. Se usó un análisis ANOVA para explorar el efecto de la enfermedad esquizofrenia, fumar diario actual, uso de alcohol, genero, edad mayor de 37 años y nivel de educación en la cantidad de cafeína usada por los consumidores. Para explorar el efecto del peso del cuerpo sobre la toma de cafeína, se incluyó el peso como variable en el modelo ANOVA. Como la distribución de la cantidad diaria de cafeína fue sesgada y causo heterocedasticidad en modelos ANOVA, se usó una transformación de raíz cuadrada de la cantidad diaria de cafeína como variable dependiente en los análisis ANOVA (Woodward, 1999). El test de Levene de igualdad de varianzas de error y pruebas de falta de ajuste sugirieron que la transformación de raíz cuadrada y los modelos ANOVA fueron apropiados.

6.4. Resultados

6.4.1. Descripción de la muestra

Cuando se compara los pacientes de esquizofrenia versus controles, aquellos tuvieron significativamente más fumadores $(69\,\%,\ 173/250\, v.\, 35\,\% 100/290;\ \chi^2 = 64{,}7,\ df = 1, p < 0{,}001)$ menos usuarios de alcohol actuales $(21\,\%,\ 52/250\, v.\, 59\,\% 172/250;\ \chi^2 = 82{,}0,\ df = 1,\ p < 0{,}001)$ y menos usuarios de drogas ilegales $(7\,\%, 17/250\, v.\, 12\,\%,\ 34/290;\ \chi^2 = 3{,}8,\ df = 1,\ p < 0{,}05)$. Dentro de los fumadores actuales, los pacientes de esquizofrenia tuvieron significativamente más fumadores duros (o sea, individuos que fuman más de 20 cigarrillos por día; $(60\,\%,\ 103/173\, v.\, 29\,\%,\ 29/100;\ \chi^2 = 23{,}7,\ df = 1,\ p < 0{,}001)$ y tuvieron puntajes de medias FTND más altos $(\text{media} \pm SD, 6{,}8 \pm 2{,}3\, v.\, 3{,}3 \pm 2{,}9; t = 10{,}3,\ df = 168, p < 0{,}001)$.

6.4.2. Toma de cafeína actual en esquizofrenia y grupo de control

Una comparación de pacientes de esquizofrenia versus controles mostró consumos de café similares (49 %,123/250 $v.$ 48 %, 138/290), pero significativamente prevalencias más bajas de consumo de té (0 % $v.$ 19 %, 55/290; $\chi^2 = 52{,}8, df = 1, p < 0{,}001$) y sodas cafeinadas (23 %, 58/250 $v.$ 33 %, 96/290; $\chi^2 = 6{,}5, df = 1, p < 0{,}01$) en pacientes. El consumo de al menos un tipo de bebida cafeinada fue presentada en 59 %(147/250) de los pacientes versus 70 %(204/290) de los controles ($\chi^2 = 7{,}87$; $df = 1$; $p < 0{,}005$). Esta diferencia permaneció significativa dentro de los sujetos que usaron tabaco y alcohol: 30 %(21/70) de pacientes versus 51 %(41/80) de controles ($\chi^2 = 6{,}95; df = 1; p < 0{,}008$).
Toma de cafeína actual (Tabla 1 apéndice C) fue más posible entre fumadores, usuarios de alcohol y gente con educación universitaria; y menos entre pacientes con esquizofrenia y sujetos mayores de 37 años. El efecto de educación sobre toma de cafeína no fue significativo en la regresión logística.

6.5. Toma de cafeína alta en consumidores de cafeína actuales

Una alta dosis de cafeína ($\geqq 200 mg/da$) fue reportada por 38 % de los pacientes de esquizofrenia (94/250) versus 25 % de controles (73/290) (la diferencia fue significativa $\chi^2 = 9{,}7; df = 1; p < 0{,}002$).
Dentro de los consumidores de cafeína, una dosis alta fue reportada por 64 %(94/147) de los pacientes de esquizofrenia versus 36 %(73/204) de los controles ($\chi^2 = 27{,}2, df = 1; p < 0{,}001$). Esta diferencia fue significativa entre los consumidores de cafeína que no usaban tabaco ni alcohol, 38 %(8/21) $v.$ 17 %(7/41)($\chi^2 = 3{,}3; df = 1; p = 0{,}07$). Entre los sujetos que actualmente consumen cafeína, una alta dosis (Tabla 2 apéndice C) fue más posible entre sujetos con esquizofrenia participantes mayores de 37 años y fumadores. Después de ajustar status y edad de los pacientes, hubo una significativa interacción entre fumar y uso de alcohol en los 351 consumidores de cafeína. La proporción de una dosis alta fue 24 %(15/62) en aquellos que no consumen ni tabaco ni alcohol, 73 %(78/107) en los que usan tabaco pero no alcohol, 52 %(47/91) en estos que usan ambos tabaco y alcohol, y 30 %(27/91) en los usuarios de alcohol pero no tabaco (comparando estos cuatro grupos, ($\chi^2 = 53{,}4; df = 3; p < 0{,}001$). Las proporciones en el primero y último grupo no fueron significativamente diferentes, pero el resto de las comparaciones a pares fue significativo, todos con $p \leqslant 0{,}003$. El patrón de esta interacción fue similar en pacientes de esquizofrenia y controles pero significativo solo en el grupo de control (ver figura 1 apéndice C).

6.6. Asociación entre toma de cafeína y otras variables

Entre los consumidores de cafeína, la cantidad o dosis fue significativamente más alta en los pacientes de esquizofrenia (media $= 200 mg/da$; $25th$ y $75th$ percentiles, 100 y 300)
versus controles (media $= 110 mg/da$; $25th$ y $75th$ percentiles, 68 y 207)
(Kruskal-Wallis$\chi^2 = 31{,}5; df = 1; p < 0{,}001$).
Entre los consumidores de cafeína después de estratificar por status de fumador, la media $\pm$ desviación estándar de la toma diaria fue $316 \pm 320 mg/da$($4{,}1 \pm 4{,}0 mg/kg/da$) en 120 fumadores con esquizofrenia, $205 \pm 157 mg/da$ ($3{,}1 \pm 2{,}5 mg/kg/da$) en 78 fumadores de control, $146 \pm 200 mg/da$($1{,}8 \pm 2{,}4 mg/kg/da$) en 27 no fumadores esquizofrénico, y $11682 mg/da$($1{,}8 \pm 1{,}3 mg/kg/da$) en 126 controles no fumadores. La toma de cafeína medida en mg/día fue significativamente diferente en los cuatro grupos (Kruskal-Wallis$\chi^2 = 72{,}9; df = 3; p < 0{,}001$).
Entre los consumidores de cafeína un análisis ANOVA de cantidad diaria de cafeína medida en mg/día sugirió efectos no significativos para el peso, género y nivel de educación. Las variables significativas fueron las mismas de las de la regresión logística descrita en la Tabla 2 apéndice C, e incluyeron esquizofrenia ($F = 9{,}5; df = 1, 345; p = 0{,}002$), edad > 37 años ($F = 6{,}1; df = 1, 345; p = 0{,}01$), fumar diario actual ($F = 41{,}1; df = 1, 345; p < 0{,}001$), y una interacción entre fumar diario y uso de alcohol ($F = 5{,}7; df = 1, 345; p = 0{,}02$). El análisis ANOVA de cantidad diaria de cafeína medida en $mg/kg/da$ en los consumidores de cafeína dio resultados similares al análisis ANOVA de la toma de cafeína medida en mg/da.
Para evaluar la asociación entre cantidad diaria de cafeína y cigarrillos fumados se calculó una correlación de Spearman (r_s) entre aquellos que consumían cafeína y cigarrillos. La correlación fue significativamente en pacientes de esquizofrenia ($r_s = 0{,}25; p = 0{,}007; n = 120$) pero no en los controles ($r_s = 0{,}12; p = 0{,}3; n = 78$). La correlación entre toma de cafeína y puntajes FTND no fue significativa (pacientes de esquizofrenia $r_s = 0{,}17, p = 0{,}07; controles, r_s =$

0,18, $p = 0{,}12$). Entre consumidores de cafeína y alcohol, no hubo asociación significativa entre cantidad de cafeína (mg/día) y alcohol (g/día) en los pacientes de esquizofrenia ($r_s = -0{,}10, p = 0{,}5; n = 47$),aunque la asociación fue significativa en controles ($r_s = 0{,}24; p = 0{,}005; n = 135$).

6.7. Discusión

Cafeína, un antagonista competitivo de receptor de adenosina, promueve la secreción de dopamina, noradrenalina, serotonina, acetylcholine, GABA glutamato (Donovan y De Vane, 2001). La cafeína acelera el procesamiento porcentual, reduce la influencia distractora de información irrelevante, y parece ejercer sus efectos más pronunciados en situaciones de fatiga (Lorist y Tops, 2003), pero una neurotransmisión dopaminergica intacta es necesaria para que la cafeína sea estimulante (Ferré, 1997). Produce efectos placenteros en dosis bajas pero no placenteros en dosis altas (Daily y Fredholm, 1998). Algunos artículos han sugerido que agonistas de adenosina selectivos pueden jugar un papel o rol como terapia adjunta para la esquizofrenia (Ferré, 1997; Dixon et al., 1999), pero también (Missak, 1991) que una sustancia tipo cafeína endógena puede ser deficiente en pacientes de esquizofrenia.
Si se propone como hipótesis una asociación fuerte entre esquizofrenia y toma de cafeína después de controlar factores confundentes, se esperaría más consumidores de cafeína entre los pacientes de esquizofrenia que en los controles y dentro de los consumidores de cafeína consumidores más «duros» entre los pacientes de esquizofrenia. Sin embargo, después de reclutar controles, hemos hallado que la esquizofrenia es asociada con menos consumidores de cafeína, pero los pacientes de esquizofrenia que consumen cafeína son más proclives a consumir altas dosis de cafeína que los controles que consumen cafeína.

6.7.1. Menos consumidores de cafeína

La menor proporción de consumidores de cafeína entre los pacientes de esquizofrenia debe ser replicada en otros estudios de control-caso. La diferencia parece ser más evidente para té (0 % en pacientes de esquizofrenia versus 19 % en controles). En España, el té no se usa ampliamente y la gente que se considera a sí misma «sofisticada» tiende a usarla más. Se sospecha que los pacientes de esquizofrenia que viven en ambientes poco «sofisticados» pueden preferir café o bebidas suaves, pero no se tienen datos para soportar esta impresión. Otra posible explicación es que los pacientes de esquizofrenia prefieren café con concentración de cafeína más alta que el té.
Después de controlar factores de confusión potenciales (genero, edad, fumar y uso de alcohol) los pacientes de esquizofrenia tuvieron aproximadamente el mismo chance ($OR = 0{,}53$) o no de consumir cafeína (Tabla 1 apéndice C). Asumiendo que este resultado sea replicado, la tasa más baja de toma de cafeína actual tiene que deberse necesariamente a una iniciación decrecida de toma de cafeína regular, a un aumento en la descontinuación de cafeína, o ambos. No se tiene información reunida sobre estos asuntos. Basados en la experiencia clínica con pacientes de esquizofrenia Españoles, la discontinuación de toma de cafeína puede estar asociada con presión de la familia sobre los pacientes que viven con sus familias. Otra posible explicación es sugerida por la observación de la cafeína que puede causar síntomas de ansiedad en algunos individuos sin esquizofrenia (Broderick y Benjamin, 2004). Es posible que los individuos vulnerables a síntomas de ansiedad inducida por cafeína pueden estar sobrerrepresentados entre los pacientes de esquizofrenia a causa de su enfermedad o medicaciones (algunos de ellos metabolizados por CYP1A2). Sin embargo, en un estudio menor de 13 pacientes de esquizofrenia, la administración aguda de cafeína bajo condiciones placebo controladas doble-ciegas no aumentó las tasas de ansiedad (Lucas et al., 1990)

6.8. Consumidores pesados de cafeína

Entre los consumidores de cafeína, el alto consumo de cafeína fue significativamente más frecuente en nuestros pacientes de esquizofrenia que en el grupo de control en análisis univariado (OR=3.2 tabla 2 apéndice C). Después de controlar los factores de confusión potenciales, esta asociación permaneció significativa con OR cerca de 2 (OR=1.9). Esta disminución en el cociente (de $OR = 3{,}2\, v. OR = 1{,}9$) parece sugerir que aunque la esquizofrenia puede tener un efecto por sí misma, parte de la asociación entre el uso pesado de cafeína y la esquizofrenia se puede explicar por algunos factores de confusión, particularmente fumar, que está asociado con la esquizofrenia. Para probar definitivamente que la esquizofrenia en sí misma es asociada con toma pesada de cafeína, es necesario controlar los efectos fármaco-kinetico de fumar midiendo los niveles de plasma de cafeína.

6.9. Asociación de toma de cafeína con nicotina y uso de alcohol

Como no era inesperado, la toma de cafeína actual fue asociada con fumar $(OR = 2{,}0)$ y uso de alcohol $(OR = 3{,}8)$ (Tabla 1 apéndice C). Cuando se estudió la cantidad de toma de cafeína en los consumidores, los resultados fueron más complicados. En el análisis univariado (Tabla 2 apéndice C) el fumar fue fuertemente asociado con toma de cafeína alta $(OR = 4{,}5)$, pero al alcohol parece decrecer el riesgo de toma de cafeína alta $(OR = 0{,}56)$ casi a la mitad. Cuando se exploró la asociación con regresión logística, se halló que el fumar tiene o tenía un efecto poderoso de aumentar el riesgo de toma de cafeína alta $(OR = 7{,}1)$, los efectos del uso del alcohol ya no fueron significativos (pero $OR > 1$) y la interacción entre fumar y uso de alcohol disminuyo el riesgo $(OR = 0{,}28)$. Esto se puede ver en la figura 1 apéndice C y significa que la toma de cafeína alta es menor en aquellos que consumen alcohol y tabaco que en aquellos consumidores de solo tabaco.

Como faltan datos sobre el curso longitudinal de fumar y toma de cafeína, es difícil discernir la sucesión cronológica de la asociación entre fumar e iniciación de toma de cafeína. Algunas interacciones fármaco-dinámicas entre cafeína y nicotina han sido descritas (Tanda y Goldberg, 2000). En ratas, el consumo crónico de cafeína acelera la adquisición de auto administración de nicotina, y la exclusión de cafeína del agua potable de animales mantenidos con nicotina resulta en una reducción de respuesta dramática, durante la primera sección libre de cafeína (Shoaib et al., 1999). También en humanos, hay alguna evidencia de que la exposición a cafeína puede potenciar los efectos de refuerzo de la nicotina (Tanda y Goldberg, 2000).

La asociación entre fumar y toma de cafeína (Tabla 2 apéndice C) se explica probablemente, al menos en parte, por un efecto fármaco-kinetico. Hidrocarburos aromáticos poli cíclicos, hallados en el humo de tabaco, inducen el cytochrome P450 1A2 (CYP1A2), y concentraciones de cafeína plasma son dos o tres veces más bajas en fumadores que en no fumadores con la misma dosis de cafeína (De Leon, 2003). Este efecto farmacocinética se ve más claramente en los pacientes de esquizofrenia que en los controles. Esto es soportado por una correlación significativa en pacientes de esquizofrenia lo cual sugiere que a más fuman, más consumen cafeína. Además, nuestro análisis previo sugiere un efecto de dosis-relación entre fumar y toma de cafeína en cantidad media, un efecto que aparece más relacionado al número de cigarrillos fumado diariamente que a la dependencia de nicotina.

Después de realizar estudios epidemiológicos en la población en general y estudios experimentales, Swanson (1994) sugirieron que el efecto farmacológico de la cafeína puede ser parcial pero no totalmente responsable de la relación entre el consumo de café y el fumar.

La asociación entre uso de alcohol y esquizofrenia parece más compleja. En un estudio suizo que comparo los pacientes de esquizofrenia con la población general, los pacientes tomaron en el cuestionario CAGE indicando que los pacientes de esquizofrenia pueden tener más problemas con el tomar alcohol que la población en general (Etter y Etter,2004).

También la asociación entre alcohol y consumo de café parece más complejo en el estudio actual. El uso de alcohol fue asociado con una mayor probabilidad de consumir cafeína, pero en los individuos que eran consumidores de cafeína, el alcohol parece disminuir su necesidad de tomar altas dosis de cafeína si ellos eran fumadores. Según la literatura, los pacientes con historia presente y pasada de abuso/dependencia de alcohol reportan altas dosis de cafeína.

La interacción compleja entre toma de cafeína, fumar tabaco y uso de alcohol no ha sido estudiada apropiadamente en estudios investigando las tres sustancias simultáneamente (Istvan y Matarazzo,1984). En la población general, el uso de cafeína parece estar débilmente asociado con el tomar alcohol, pero fuertemente asociada con el fumar tabaco Esta relación triple entre cafeína, alcohol y cigarrillo puede ser explicada al menos en parte por factores genéticos (Hettema y otros 1999). En un estudio de alcohólicos Israelíes Amit y otros (2004) hallaron que, en el subgrupo de sujetos con una historia familiar de alcoholismo, Hubo asociaciones entre toma de alcohol y uso de cafeína entre toma de alcohol y fumar y entre cafeína y fumar. En sujetos sin historia familiar de alcoholismo se halló relaciones entre alcohol, cafeína y fumar, sin embargo, la toma de café y el tabaco no fueron relacionadas. Amet y otros(2004) sugieren que sus resultados parecen ser consistentes con una noción de interacción entre estos comportamientos de usos de sustancias, que pueden ocurrir de modo comportamental más que genético. La impulsividad que es más alta en pacientes usuarios de substancias podría ser un factor medianamente común, y el abuso parece ocurrir entre el prodromo y el primer episodio psicótico. Si embargo, aunque los fumadores pueden no mostrar alta impulsividad, si pueden mostrar una alta desinhibición.

6.10. La comparación de la asociación entre la esquizofrenia y la cafeína versus con la esquizofrenia y el tabaquismo

Existe una fuerte asociación entre esquizofrenia y el tabaquismo, que se ha observado en muchos estudios (de Leon y Díaz, 2005). Cuando se comparan las muestras de esquizofrénicos con los controles se encontró: 1) la esquizofrenia se asoció significativamente con el tabaquismo (Gurpegui et al., 2005); 2) la esquizofrenia se asoció significativamente con el tabaquismo en los fumadores pesados (Gurpegui et al., 2005; y 3) Dentro de los pacientes con esquizofrenia, algunos síntomas esquizofrénicos y los resultados a largo plazo fueron significativamente asociados con la dependencia de la nicotina (Aguilar et al., 2005). Con respecto a la asociación entre el consumo de cafeína y la esquizofrenia, se encontró que: 1) la esquizofrenia se asoció significativamente con el consumo actual menos frecuente de cafeína; 2) esquizofrenia se asoció significativamente con un mayor consumo de cafeína entre los consumidores de cafeína; y 3) síntomas de la esquizofrenia y los resultados a largo plazo no se asociaron significativamente con el consumo de cafeína (Gurpegui et al., 2004).

Las asociaciones mencionadas anteriormente entre el consumo de cafeína y la esquizofrenia no parecen ser tan fuerte y consistente como la asociación entre fumar y la esquizofrenia (Gurpegui et al., 2005). Por otra parte, una vulnerabilidad a la esquizofrenia puede estar asociada con un mayor riesgo de convertirse en un fumador diario; lo que fue apoyado por nuestra muestra de pacientes, ya que las tasas de tabaquismo de iniciación fueron mayores en los pacientes con esquizofrenia que comenzaron a fumar a diario por lo menos cinco años antes de inicio de la enfermedad (Gurpegui et al., 2005)

El estudio de Gurpegui et al. (2004) es el único estudio de cafeína en la literatura que controla los factores de confusión, y la adición de la muestra de control ha hecho posible una mayor precisión en la comparación de control de casos. Sin embargo, el actual estudio estaba limitado por la falta de mediciones de las concentraciones plasmáticas de cafeína. En estudios futuros, la medición de las concentraciones de cafeína proporcionará una mejor definición de la ingesta fuerte de cafeína y un control de los efectos farmacocinéticos de fumar. Del mismo modo, en futuros estudios, la evaluación de los problemas de alcohol, usando el CAGE o de otros instrumentos, ayudará a evaluar la ingesta de alcohol de una manera más sofisticada.

En conclusión, este estudio, que necesita replicación, sugirió que el consumo diario de cafeína fue menos frecuente en los pacientes con esquizofrenia $(59\,\%, 147/250)$ que en los controles $(70\,\%, 204/290)$. Sin embargo, entre los consumidores de cafeína, una ingesta abundante de cafeína $(\geqslant 200 mg/da)$ fue significativamente más frecuente en los pacientes con esquizofrenia $(64\,\%, 94/147)$ que en el control $(36\,\%, 73/204)$. Los análisis multivariados sugieren que estas diferencias no pueden ser explicadas por factores de confusión.

Apéndices

Apéndice A

Ejemplo 4.3.1 Regresión logística con una variable binaria (crea los Cuadros 4.2 y 4.3)

```
datapooling;
/*read in the data */
input smoking r n;
cards;
1 166 1342
0 50 563
;
/* fit a logistic regression model with explanatory
variable SMOKING*/
proc genmod;
/* include variance-covariance matrix using COVB*/
model r/n = smoking/dist=binomial link= logit covb;
run;
```

1) Modelos de factor común: se tiene una vulnerabilidad compartida (el rango de vulnerabilidades abarca de lo genético a lo sociocultural para ambos tipos de desórdenes);
2) Modelos de adicción secundaria: la adicción es el resultado de SMI (este incluye el modelo de auto-medicamento);
3) Modelos SMI secundario: SMI es el resultado de adicción;
4) Modelos Bidireccionales: combina el segundo y tercer tipo de modelos

Cuadro A.1: Modelos empleados para explicar la asociación entre adicción y desórdenes mentales severos (SMIs) (Mueser et al. 1998.)

ARGUMENTOS CONTRA EL EMPLEO DE LA POBLACIÓN GENERAL COMO CONTROLES 1)
-Los Pacientes con esquizofrenia son muy diferentes de las personas de la población general.
-Los pacientes con esquizofrenia tienden a residir en lugares no-standard's.
-Los pacientes con esquizofrenia no son propensos a participar en cirugías. Durante loa últimos 30 años, las agencias guvernamentales de US han conducido 5 conjuntos diferentes de estudios epidemiológicos (Ross 2008) no ha proporcionado casi datos sobre esquizofrenia.[1]
ARGUMENTOS EN FAVOR DEL EMPLEO DE SMIs COMO CONTROLES
SMIs son usualmente asociados con factores que pueden contribuir al fumar ó al fumar en exceso tales como: 1) nivel de educación inferior en los US ó niveles económicos bajos en otros países,[2] 2) exposición a entornos con niveles superiores consumo de cigarrillo tales como hospitales psiquíatricos y clínicas, y 3) consumo de medicamentos psiquíatricos.[3]

[1] En 1980, el primer estudio, el Epidemiological Catchment Area (ECA), hizo provisiones para reclutar poblaciones institucionalizadas incluyendo pacientes esquizofrénicos. Desafortunadamente, este estudio no recolectó datos sobre el consumo de cigarrillo excepto de algunos centros (Covey et al. 1994.) que no exploran las asociaciones entre esquizofrenia y fumar. Similarmente, la National Comorbidity Survey (NCS) no proporciona datos específicos sobre esquizofrenia y fumar (Lasser et al. 2000.) La National Epidemiology Survey on Alcohol and Related Conditions (NESARC) describió dependencia de nicotina, pero no incluyó pacientes esquizofrénicos y fumadores (Grant et al. 2004). Más recientemente, en el 2007 la National Health Survey (NHS) (McClave el al. 2010.) no incluyeron pacientes institucionales excepto 150 pacientes con esquizofrenia; las edades de prevalencia para fumadores varían entre entre 59,1 % en comparación de un 18,3 % en adultos no-psiquíatricos reportados. Este estudio no considera la depresión como un factor de confusión perteneciente a pacientes no-psiquíatricos y no es correcto para las diferencias de género.

[2] Los pacientes con esquizofrenia tienden a tener recursos económicos más bajos que la población general; los otros pacientes con SMIs pueden ser mejor controlados desde la accesibilidad y disponibilidad para una influencia de drogas para el empleo de las misma y la adicción (Anderson 2006.).

[3] En pacientes con esquizofrenia, algunos antipsicóticos de primera generación, tales como haloperidol, pueden contribuir al consumo excesivo de cigarrillo de acuerdo a dos estudios abiertos (McEvoy et al. 1995; Kim et al. 2010), y a un cese en ambos casos clínicos (George et al. 2000) y al surgimiento del vegetarianismo (Gonzalez-Pinto et al 2011). Los efectos negativos de los antipsicóticos de primera generación sobre el fumar es una pobibilidad, pero no se han demostrados casos de estudio sobre este tema y son pequeños y limitados en diseño (Matthews et al. 2011.)

Cuadro A.2: Controles de estudio: argumentos contra el empleo de la población general y en favor del uso de SMIs.

Cuadro: A.3. Resumen de meta-análisis sobre la asociación entre esquizofrenia y diferentes comportamientos del fumar: odds ratios (ORs) y consistencia

	Estudio de consistencia[a]			Países		Estimación ORs (95% CI)		
	N	OR>1	+ Significativo	N^b	Clase[c]	Ejemplo total	Hombres	Mujeres
↑ Fumadores corrientes vs población general	42	40/42	37/42	20	Algunos no-Occidentales	5.3(4.9-5.7)		
	32	29/32	29/32	18	Mas occidentales		7.2 (6.1-8.3)	
	25	20/25	20/25	15	Mas occidentales			3.3 (3.0-3.6)
↑ Fumadores corrientes vs otros SMIs	18	17/18	14/18	9	Mas occidentales	1.9 (1.7-2.1)		
	14	11/14	11/14	8	Mas occidentales		2.3 (2.0-2.7)	
	8	6/8	3/8	5	Occidentales			1.8 (1.5-2.3)
↑ Fumadores ocasionales vs población general	9	8/9	8/9	6	Occidentales	3.1 (2.4-3.8)		
	4	4/4	3/4	3	Occidentales		7.3 (1.04-13.6)	
	3	3/3	2/3	3	Occidentales			2.8 (1.2-4.4)
↑ Fumadores ocasionales vs otros SMIs	5	4/5	3/5	5	Mas occidentales	2.0 (1.6-2.4)		
	4	4/4	2/4	3	Mas occidentales		2.0 (1.5-2.7)	
	2	1/2	0/2	2	Occidentales			0.92 (0.44-1.9)
↓ Dejar de fumar en fumadores ocasionales vs población general	6	6/6	6/6	5	Occidentales	5.3 (4.2-7.1)		
	4	4/4	4/4	3	Occidentales		10.0 (7.1-16.7)	
	3	3/3	2/3	3	Occidentales			2.2 (1.5-4.3)
↓ Dejar de fumar en fumadores ocasionales vs otros SMIs	4	3/4	1/4	4	Occidentales	1.8 (1.1-3.0)		
	4	4/4	3/4	3	Mas occidentales		2.9 (1.9-13.4)	
	2	0/2	0/2	2	Occidentales			0.37 (0.11-1.3)

CI:95% intervalo de confianza

Consistencia[a]: incluída N: número de estudios. RO> 1:número de estudios individuales con $OR > 1$ dividido por N. +Significativo: número de estudios individuales con $OR > 1$significativo dividido por N

N[b]:número de países

Clase[c]: USA, Canadá, Europa e Israel son considerados países Occidentales

Apéndice B

Smoking initation and schizophrenia: a replication study in Spanish sample

Apéndice C

Fewer but caffeine consumers in schizophrenia: A case-control study

Bibliografía

Agrawal, A, JL Silberg, MT Lynskey, and LJ Evans. "Mechanisms underlying the lifetime co-occurrence of tobacco and cannabis use in adolescent and young adult twins." *Drug Alcohol Depend*, no. 108:49–55.

Aguilar, MC, M Gurpegui, FJ Diaz, and de J Leon. 2005. "Nicotine dependence and symptoms in schizophrenia: naturalistic study of complex interactions." *Br. J. of Psychiatry*, no. 186:125–221.

Anderson, P. "Global use of alcohol, drugs and tobacco." *Drug Alcohol Rev*, no. 25:489–502.

Association, American Psychiatric. 1980. *Diagnostic and statistical manual*. third DSM-III. American Psychiatric Press.

Barnes, M, BR Lawford, SC Burton, KR Heslop, K Hausdorf, and RM Young. "Smoking and schizophenia: is symptom profile related to smoking and which antipsychotic is of benefit in reducing cigarrete use?" *Aust NZJ Psychitry*, no. 40:575–580.

Berrios, GE, R Luque, and JM Villagran. "Schizophrenia: a conceptual history." *Int. J. Psychology*, no. 3:111–140.

Biedermann, F, and WW Fleischhacker. 2011. "Emerging drugs for schizophrenia." *Expert Opin Emerg Drugs*, no. 16:271–282.

Bien, TH, and R Burge. 1990. "Smoking and drinking: a review of the literature." *Int. J. Addict*, no. 25:1429–1454.

Bierut, LJ. 2011. "Genetic vulnerability and susceptibility to substance dependence." *Neuron*, no. 69:618–627.

Black, DW, M Zimmerman, and WH Coryell. 1999. "Cigarette smoking and psychiatric disorder in community sample." *Ann. Clin. Psychiatry*, no. 11:129–136.

Bleuler, D. 1987. *Dementia praecox or the group of schizophrenias*. International University Press, Madison, CT.

Boardman, JD, CL Blalock, and FC Pampel. 2010. "Trends in the genetic influences on smoking." *J. Health Soc. Behav.*, no. 51:108–123.

Bolton, JM, and J Robinson. 2010. "Population-attributable fractions of Axis I and Axis II mental disorders for suicide attempts: findings from a representative sample of the adult, noninstitutionalized US population." *Am. J. Public Heath*, no. 100:2473–2480.

Bozikas, VP, I Niopas, A Kafantari, FI Kanaze, C Gabrieli, P Melissidis, K Gamvrula, K Fokas, and A Karavatos. 2005. "No increased levels of the nicotine metabolic cotinine in smokers with schizophrenia." *Prog. Neuropsychopharmacol Biol. Psychiatry*, no. 29:1–6.

Breslau, N, SP Novak, and RC Kessler. 2004a. "Daily smoking and the subsequent onset of psychiatric disorders." *Psycol. Med.*, no. 34:323–333.

———. 2004b. "Psychiatric disorders and stages of smoking." *Biol. Psychiatry*, no. 55:69–76.

Buckley, PF. 2006. "Prevalence and consequences of the dual diagnosis of substance abuse and several mental illness." *J. Clin. Psychiatry*, no. 67 (suppl. 7):5–9.

Buckley, PF, BJ Miller, DS Lehrer, and DJ Castle. 2009. "Psychiatric comorbidities and schizophrenia." *Schizophr. Bull.*, no. 35:383–402.

Budney, AJ, ST Higgins, and JR Hughes WK Bickel. 1993. "Nicotine and caffeine use in cocaine-dependent individuals." *J. Subst. Abuse*, no. 5:117–130.

Campo, A, LA Díaz, GE Rueda, M Rueda, and D Farelo. 2004. "Prevalencia de consumo de cigarillo en pacientes de la consulta psiquiátrica de Bucaramanga." *Colombia Médica*, no. 35:69–74.

Campo-Arias, A, LA Díaz-Martínez, Rueda-Jaimes, M Rueda-Sánchez, D Farelo-Palacín, FJ Díaz, and de J Léon. 2006. "Smoking is associated with schizophrenia, but not with mood disorders, within a population with low smoking rates: a matched case-control study in Bucaramanga, Colombia." *Schizophr. Res.*, no. 83:269–276.

Chapman, S, M Ragg, and K McGeechan. 2009. "Citation bias in reported smoking prevalence in people with schizophrenia." *Aust. N Z J Psychiatry*, no. 43:277–282.

Covey, LS, AH Glassman, and F Stetner. 1998. "Cigarette smoking and major depresion." *J. Addict Dis*, no. 17:35–46.

Covey, LS, DC Hughes, AH Glassman, DG Blazer, and LK George. 1994. "Ever-smoking, quitting, and psychiatric

disorders: evidence from the Durham, North Carolina and Epidemiological Catchment Area." *Tob. Control*, no. 3:222–227.

Díaz, FJ, D James, S Botts, L Maw, and MT Susce J de León. 2009. "Tobacco smoking behaviors in bipolar disorder: a comparison with the general population, schizophrenia and major depresion." *Bipolar Disord.*, no. 11:154–165.

Díaz, FJ, M Jané, E Saltó, H Pardell L Salleras, C Pinet, and J de León. 2005. "A brief measure of high nicotine dependence for busy clinicians and large epidemiological surveys." *Aust. N Z J Psychiatry*, no. 39:161–168.

Díaz, FJ, DM Velasquez, and MT Susce J de León. 2008. "The association between schizophrenia ans smoking: Unexplained by either the illness of prodromal period." *Schizophr. Res.*, no. 104:214–219.

Degenhardt, L, WT Chiu, N Sampson, RC Kessler, JC Anthony, M Angermeyer, R Bruffaerts G Girolomano, O Gureje, Y Huang A Karam S Kostyuchenko, JP Lepine, ME Mora, Y Neumark, JH Ormel, A Pinto-Meza, and J Posada-Vil. 2008. "Toward a global view of alcohol, tabacco, cannabis, and cocaine use: findings from the WHO World Mental Health Surveys." *PLoS Med.* 5, no. e141.

Degenhardt, L, L Dierker, WT Chiu, ME Medina-Mora, Y Neumark, N Sampson, J Alonso, M Angermeyer, JC Anthony, R Bruffaerts, G de Girolamo, R de Graaf, O Gureje, AN Karam, S Kostyuchenko, S Lee, and JP Lépine. 2010. "Evaluating the drug use «gateway» theory using cross-national data: consistency and associations of the order of initiation of drug use among participants in the WHO World Mental Health Surveys." *Drug. Alcohol Depend*, no. 108:84–97.

Degenhardt, L, and W Hall. 2001a. "The relationship between tobacco use, substance-use disorders and mental health: results from the National Survey of Mental Health and Well-Being." *Psychol. Med.*, no. 3:225–234.

———. 2001b. "The association between psychosis and problematical drug use among Australian adults: findings from the National Survey of Mental Health and Well-Being." *Psychol. Med.*, no. 31:659–668.

Degenhardt, L, W Hall, and M Lynskey. 2001. "Alcohol, cannabis and tobacco use among Australians: a comparison of thier associations with other drug use and use disorders, affective and anxiety disorders, and psychosis." *Addiction*, no. 96:1603–1614.

de León. 1996. "Smoking and vulnerability for schizophrenia." *Schizophr. Bull*, no. 22:405–409.

de León, J, E Becoña, M Gurpegui, A Gonzalez-Pinto, and FJ Díaz. 2002b. "The association between high nicotine dependence and severe mental illness may be consistent across countries." *J. Clin. Psychiatry*, no. 63:812–816.

de León, J, and F Díaz. 2005. "A meta-analysis of worldwide studies demonstrates an association between schizophrenia and tobacco smoking behaviors." *Schizophr. Res.*, no. 76:135–157.

de León, J, FJ Díaz, MC Aguilar, D Jurado, and M Gurpegui. 2006. "Does smoking reduce akathisia?: Testing a narrow version of the self-medication hypothesis." *Schizophr. Res.*, no. 86:256–268.

de León, J, FJ Díaz, E Becoña, M Gurpegui, D Jurado, and A Gonzalez-Pinto. 2003. "Exploring brief measures of nicotine dependence for epidemiological surveys." *Addict. Behav.*, no. 28:1481–1486.

de León, J, FJ Díaz, T Rogers, D Browne, and L Dinsmore. 2002a. "Initiation of daily smoking and nicotine dependence in schizophrenia and mood disorders." *Schizophr. Res.*, no. 56:47–54.

de León, J, M Gurpegui, and FJ Díaz. 2007a. "Epidemiology of comorbid tobacco use and schizophrenia: thinking about risk and protective factors." *J, Dual Diagn*, no. 3:9–25.

de León, J, DM Rendón, E Baca-García, F Aizpuru, A Gonzalez-Pinto, C Anitua, and FJ Díaz. 2007b. "Association between smoking and alcohol use in the general population: Stable and unstable odds rations across two years in two different countries." *Alcohol*, no. 42:252–257.

de León, J, MT Susce, FJ Díaz, DM Rendón, and DM Velásquez. 2005. "Variables associated with alcohol, drug and daily smoking cessation in patients with severe mental illnesses." *J. Clin. Psychiatry*, no. 66:1447–1455.

de León, J, J Tracy, E McCann, A McGrory, and FJ Díaz. 2002c. "Schizophrenia and tobacco smoking: A replication study in another US psychiatric hospital." *Schizophr. Res.*, no. 56:55–65.

Dick, DM, JL Meyers, RJ Rose, J Kaprio, and KS Kendler. 2011. "Measure of current alcohol consumption and problems:two independent twin studies suggest a complex genetic architecture." *Alcohol Clin. Exp. Res.*, Jun 20 [Epub ahead of print].

Dierker, L, J He, A Kalaydijan, J Swendsen, L Degenhardt, M Glantz, K Conway, J Anthony, WT Chiu, NA Sampson, R Kessler, and K Merikangas. 2008. "The importance of timing of transitions for risk of regular smoking and nicotine dependence." *Ann. Behv. Med.*, no. 36:87–92.

Dobson, Annette L. 1990. *An Introduction to Generalized Linear Models*. Chapman & Hall/CRC.

Dome, P, Z Rihmer, X Gonda, HG Kiss, D Kovács, K Seregi, and Z Teleki. 2005. "Cigarette smoking and psychiatric disorders in Hungary." *Neueopsychopharmacol Hung.*, no. 9:145–148.

Edwards, AC, and KS Kendler. 2011. "Nicotine withdrawal-induced negative affect is a funtion of nicotine dependence and not liability to depression or anxiety." *Nicotine Tob Res.*, no. 13:677–685.

Edwards, AC, HH Maes, NL Pedersen, and KS Kendler. 2011. "A population-based twin study of the genetic and environmental relationship of major depression, regular tobacco use and nicotine dependence." *Psychol. Med.*, no. 41:395–405.

Esterberg, ML, EM JOnes, MT Compton, and EF Walker. 2007. "Nicotine consumption and schizotypy in fist-degree relatives of individuals with schizophenia and non-psychiatric controls." *Schizophr. Res.*, no. 97:6–13.

Falk, DE, HY Yi, and S Hiller-Sturtmhöfel. 2006. "An epidemiologic analysis of co-occurring alcohol and tobacco use and disorders: findings from the National Epidemiologic Survey on Alcohol and Related Conditions." *Alcohol Res. Health.*, no. 29:162–171.

Fanous, AH, MC Neale, CO Gardner, BT Webb, RE Straub, FA OÑeill, D Walsh, BP Riley, and KS Kendler. 2007. "Significant correlation in linkage signals from genome-wide scans of schizophrenia and schizotypy." *Mol. Psychiatry.*, no. 12:958–965.

Freedman, R, H Coon, M Myles-Worsley, A Orr-Urtreger, A Olincy, A Davis, M Polymerpoulos, J Holik, J Hopkins, M Hoff, J Rosenthal, MC Waldo, F Reimherr, P Wender, J Yaw, DA Young, CR Breese, and C Adams a. 1997. "Linkage of a neurophysiological deficit in schizophenia to a chromosome 15 locus." *Proc. Nat. Acad. Sci. USA.*, no. 587-592.

Frees, Edward W. 2004. *Longitudinal and Panel Data*. Cambridge University Press.

Gejman, PV, AR Sanders, and KS Kendler. 2011. "Genetic of schizophrenia: new findings and challenges." *Annu. Rev. Genomics. Hum. Genet.*, no. 12:121–144.

George, TP, DM Ziedonis, A Feingold, WT Pepper, CA Satterburg, J Winkel, BJ Rounsaville, and TR Kosten. 2000. "Nicotine transdermal patch and atypical antipsychotic medications for smoking cessation in schizophrenia." *Am. J. Psychiatry.*, no. 157:1835–1842.

Gershon, Grand RB, S Hwang, J Han, T George, and AL Brody. 2007. "Short-term naturalistic treatment outcomes in cigarette smokers with substance abuse and/or mental illness." *J. Clin. Psychiatry.*, no. 68:892–898.

González-Pinto, A, S Alberich, S Ruíz de Azúa, M Martínez-Cengotitabengoa, M Fernández, M Gutiérrez, M Saenz, A Besga, P Galdós, and J de León. 2011 Agu 31 [Epub ahead of print]. "Psychosis and smoking cessation: difficulties in quitting associated with sex and substance abuse." *Psychiatry. Res.*

Goodman, A. 1990. "Addiction: definition and implications." *Br. J. Addict.*, no. 85:1403–1408.

Grant, BF, DS Hasin, SP Chou, FS Stinson, and DA Dawson. 2004. "Nicotine dependence and psychiatric disorders un the United States: results from the national epidemiologic survey on alcohol and related conditions." *Arch. Gen. Psychiatry.*, no. 61:1107–1115.

Greenland, S, and K Drescher. 1993. "Maximium likelihood estimation of the attributable fraction from logistic models." *Biometrics.*, no. 49:865–872.

Gregg, L, Barrowclough, and G Haddock. 2007. "Reasons for increased substance use in psychosis." *Clin. Psychol. Rev.*, no. 27:494–510.

Gregg, L, C Barrowclough, and G Haddok. 2007. "Reason for increased substances use in psychosis." *Cli. Psychol. Rev.*, no. 27:494–510.

Gurpegui, M, JM Martínez-Ortega, MC Aguilar, FJ Díaz, HM Quintana, and J de León. 2005. "Smoking initiation and schizophenia: a replication study in a Spanish sample." *Schizophr. Res.*, no. 76:113–118.

Gurpegui, M, JM Martínez-Ortega, MC Aguilar, D Jurado, FJ Díaz, and J de León. 2007. "Subjective effects and the main reason for smoking in outpatients with schizophenia: a case-control study." *Compr. Psychiatry.*, no. 48:186–191.

Harris, JG, S Kong, D Allensworth, L Martin, J Tregellas, B Sullivan, G Zerbe, and R Freedman. 2004. "Effects of nicotine on cognitive deficits in schizophrenia." *Neuropsychopharcology.*, no. 29:1378–1385.

Heatherton, TF, LT Kozlowski, RC Frecker, and KO Fagerström. 1991. "The Fagerström test for nicotine dependence: A version of the Fagerström Tolerance Questionnarie." *Br. J. Addict.*, no. 86:1119–1127.

Heatherton, TF, LT Kozlowski, RC Frecker, W Rickert, and J Robinson. 1998. "Measuring the heaviness of smoking: using self-reported time to the first cigarette of the day and number of cigarettes per day." *Br. J. Addict.*, no. 84:791–800.

Henningfeld, JE, C Cohen, and J Slade. 1998. "Is nicotine more addictive tha cocaine?" *Br. J. Addict.*, no. 86:565–569.

Hilbe, Joseph M. 2009. *Logistic Regression Models*. CRC Press. Taylor & Francis Group.

Hitsman, B, B Borelli, DE McChargue, B Spring, and R Niaura. 2003. "History of depression and smoking cessation outcome: a meta-analysis." *J. Consult. Clin. Psychol.*, no. 71:657–663.

Hughes, JR. 1996. "An overview of nicotine use disorders for alcohol/drug abuse clinicians." *AJA.*, no. 5:262–273.

Hughes, JR, DK Hatsukamy, JE Mitchell, and LA Dahlgren. 1986. "Prevalence of smoking among psychiatric outpatients." *Am. J. Psychiatry.* 143, no. 993-997.

Ioannidis, JP. 2005. "Why most published research findings are false." *PLoS Med.*, no. 2:8–24.
Itkin, O, B Nemets, and H Einat. 2001. "Smoking habits in bipolar and schizophrenic outpatients in southern Israel." *J. Clin. Psychiatry.*, no. 63:368–369.
Jaspers, K. 1963. *General psychopathology.* University of Chicago Press.
Woodward, Mark. 1999. *Epidemiology. Study design and data analysis*. Chapman & Hall/CRC.