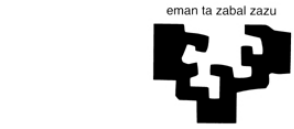

ZTF-FCT
Zientzia eta Teknologia Fakultatea
Facultad de Ciencia y Tecnología

Aplicación de metodología estadística para predecir la mortalidad de pacientes con neumonía

Trabajo Fin de Grado
Grado en Matemáticas

Laura Ansola Marlasca

Trabajo dirigido por
Irantzu Barrio Beraza

Leioa, Junio de 2015

Índice general

Resumen

La regresión logística es un caso particular del modelo lineal generalizado, en el que se relaciona la variable respuesta binomial con el resto de covariables, que pueden ser tanto continuas como categóricas. La distribución más frecuente para la variable respuesta Y es la binomial con valores 0 y 1, y la regresión logística relaciona la probabilidad de que suceda el evento Y con el resto de variables $\mathbf{X}$. Por tanto, estudiaremos en profundidad como crear esa función que modeliza la esperanza de la variable respuesta mediante una función link, y aprenderemos a interpretar los resultados.

En esta memoria vamos a estudiar los factores de riesgo que influyen en que un paciente de neumonía fallezca durante su ingreso, para así crear un score que ayude a los médicos en el proceso de la toma de decisiones. Con intención de poder determinar la gravedad del paciente casi de inmediato, crearemos dos modelos. El primero incluirá las variables cuyos resultados puedan conocerse en una consulta rutinaria de unos 15 minutos; en cambio el otro, incluirá factores de riesgo para los que se necesitan los resultados de un análisis de sangre y de una gasometría.

En primer lugar, crearemos los dos modelos que relacionan las variables que hemos seleccionado con la variable respuesta mortalidad. En base a criterio y recomendación clínica incluiremos todas las variables continuas como dicotómicas, distinguiendo entre los valores anómalos y los normales. Una vez conocemos las variables que influyen en la mortalidad del paciente, vamos a proceder a la creación de los scores. Para eso, emplearemos dos sistemas de puntuación que asignan puntos a las distintas categorías de cada factor de riesgo, uno por cada modelo. Las categorías de referencia (los valores normales, menos graves) tomarán 0 puntos en el sistema, y por tanto, el paciente que obtenga mayor puntuación en el marcador, se considerará más grave.

Una vez creados los dos marcadores, queremos ver si son buenos predictores de mortalidad al aplicarlos a nuevos individuos. Para eso llevaremos a cabo lo que se conoce como la validación de los scores. La validación se realizará mediante lo que se conoce como *split-sample validation*, método que consiste en dividir la muestra original en dos submuestras: muestra de derivación y muestra de validación. El modelo se desarrolla en la muestra de derivación,

mientras que a la muestra en la que se valida el modelo se le llama muestra de validación. Dibujaremos las curvas ROC de los dos scores en ambas muestras y compararemos el área bajo esas curvas (el AUC) en la muestra de derivación y en la de validación mediante el test de `Delong`.

Por último, vamos a comparar los resultados que obtenemos en ambos marcadores con otros dos scores que utilizan los neumólogos: el CURB-65 y el FINE. El primero, es un score muy simple que consta de tan solo 5 variables. El FINE en cambio, es un marcador mucho más complejo que requiere de pruebas específicas para poder asignarle una puntuación al paciente. Ambos scores, utilizan el método de *split-sample* para la validación del modelo.

El objetivo principal de este trabajo es proporcionar a los expertos una herramienta que les ayude a determinar la gravedad del paciente fácilmente.

Agradecimientos

Me gustaría expresar mi agradecimiento al equipo de neumólogos del Hospital Universitario de Cruces por ayudarme a entender cómo afecta una neumonía en el organismo y por explicarme en qué factores se aprecia la gravedad de la enfermedad, en especial al Dr. Rafael Zalacain Jorge por haberme facilitado los recursos necesarios para llevar a cabo esta memoria.

Capítulo 1

Regresión logística

El modelo lineal general relaciona la variable independiente continua Y con las covariables $\mathbf{X} = (X_1, ..., X_q)$ bajo la hipótesis de que la variable respuesta Y sigue una distribución normal tal que $Y : N(0, \sigma^2)$. Cuando no se cumple la hipótesis de que la variable respuesta sea normal, se puede extender este modelo al caso modelo lineal generalizado donde la esperanza de la variable respuesta se modeliza mediante una función link.

En el ámbito clínico la distribución más frecuente para el outcome de interés es la distribución binomial con valores 0 y 1. La variable respuesta Y indicará si sucede el evento Y $(Y = 1)$ o si no $(Y = 0)$. Para estudiar la relación entre las covariables y este tipo de variable respuesta de la familia exponencial se utiliza un caso particular del modelo lineal generalizado, la regresión logística.

1.1. Introducción

Empezaremos por explicar la regresión logística simple, esto es, la relación entre una sola covariable X con la variable respuesta Y. Esta covariable puede ser continua, dicotómica o categórica. Dada X, se define como $E(Y|X)$ al valor esperado de Y en función de X. Al tratarse de una variable respuesta dicotómica, Y sigue una distribución binomial, $Y : Bin(1, p)$ donde $p = P(\textit{éxito})$. Por tanto, en este caso, tendremos $E(Y|X) = p$ que tomará valores entre 0 y 1.

Tal y como hemos dicho, la regresión logística es un caso particular del modelo lineal generalizado y por tanto, la función que relaciona la esperanza de la variable respuesta con las demás covariables, también debe tomar valores en toda la recta real. Por tanto, necesitamos definir una función que convierta p en continua y que la escriba en función de X, para poder estudiar así la relación entre las dos. La probabilidad toma valores en $[0, 1]$ y la función que

vamos a definir ahora los tomará entre $(-\infty, \infty)$, igual que los coeficientes de la misma.

$$g(E(Y|X)) = \beta_0 + \beta_1 X \tag{1.1}$$

La función link que usaremos es la siguiente:

$$g(p) = logit(p) = ln\frac{p}{1-p} \tag{1.2}$$

Si partimos de una muestra con n observaciones independientes del tipo (x_i, y_i), $i = 1, ..., n$, este es el *logit* que obtenemos:

$$logit(p_i) = ln\frac{p_i}{p_i - 1} = \beta_0 + \beta_1 x_i, \tag{1.3}$$

donde $p_i = p(x_i)$ indica la probabilidad de cada individuo i.

Despejando p_i de la expresión 1.3 obtenemos:

$$p_i = \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}} \tag{1.4}$$

Supongamos ahora que tenemos q variables. Nos referiremos a estas q variables o covariables a través del vector $\mathbf{X}$= $(X_1, X_2, ..., X_q)$. Denotaremos $p(\mathbf{X})$ a la probabilidad que tiene la variable respuesta Y de tomar el valor 1 a partir de esas p variables, es decir, $p(\mathbf{X}) = P(Y = 1|\mathbf{X})$. Por tanto, siguiendo la notación anterior, el logit del modelo de regresión logística múltiple es:

$$g(p(\mathbf{X})) = ln\frac{p(\mathbf{X})}{1 - p(\mathbf{X})} = \beta_0 + \beta_1 X_1 + ... + \beta_q X_q \tag{1.5}$$

Asimismo, la probabilidad $p(\mathbf{X})$ es:

$$p(\mathbf{X}) = \frac{e^{g(p(\mathbf{x}))}}{1 + e^{g(p(\mathbf{x}))}} \tag{1.6}$$

No todas las covariables han de ser continuas, por lo que tenemos que cambiar un poco la función (1.5). En el caso de que la variable X_j tuviera más de dos categorías, con $j \in [1, q]$, no podríamos respresentar la variable X_j con un solo coeficiente β_j. Si una variable nominal consta de k_j categorías deben crearse $k_j - 1$ variables dicotómicas a las que llamaremos variables *dummies* o variables diseñadas [1]. Denotaremos estas $k_j - 1$ variables como $(D_{j1}, D_{j2}, ..., D_{jk_j-1})$. A cada categoría o clase de la variable nominal le corresponde un conjunto de valores de los D_{js} con el que se identifica dicha base, con $s = 1, ..., k_j - 1$.
La manera más usual de definir las $k_j - 1$ variables dummies es la siguiente: si el sujeto pertenece a la primera categoría, entonces todas valen 0: (D_{j1} =

$D_{j2} = ... = D_{jk_j-1} = 0$); si el sujeto se halla en la segunda categoría, ($D_{j1} = 1$ y $D_{j2} = ... = D_{jk_j-1} = 0$); y así sucesivamente hasta llegar a la última categoría, para la cual $D_{jk_j-1} = 1$ y las restantes variables dummies valen 0.

Supongamos, por ejemplo, que X_j indica si un individuo fuma, es ex-fumador o si, por el contrario, nunca ha fumado; es decir, $X_j = 0$ si no fuma, $X_j = 1$ si fuma y $X_j = 2$ si es ex-fumador. En ese caso, como la variable consta de $k_j = 3$ categorías, necesitaríamos $k_j - 1 = 2$ variables dicotómicas (0 vs. 1) para poder explicar X_j: D_{j1} y D_{j2}. Estos serían los valores de las variables dummy:

$\mathbf{X_j}$ = **Tabaquismo**	D_{j1}	D_{j2}
No fumador	0	0
Fumador	1	0
Ex-fumador	0	1

Como vemos, si el individuo nunca ha fumado, los valores de las variables dummies son cero, y por tanto los coeficientes β_{j1} y β_{j2} que los acompañan no se sumarán (o restarán) al modelo de regresión. Por eso decimos que tomamos esta categoría (normalmente $X_j = 0$) como la de referencia.

Como hemos dicho, al ajustar un modelo que incluya una variable X_j categórica con k_j niveles, ésta debe ser sustituida por las $k_j - 1$ variables dummies D_{js}, y a cada una de ellas le corresponderá su respectivo coeficiente β_{js}, con $s = 1, ..., k_j - 1$. Por tanto, este será el logit para un modelo con q variables siendo la variable j categórica y de k_j niveles:

$$g(p(\mathbf{X})) = \beta_0 + \beta_1 X_1 + ... + \sum_{s=1}^{k_j-1} \beta_{js} D_{js} + ... + \beta_q X_q \tag{1.7}$$

1.1.1. Estimación de los coeficientes

Para estimar los coeficientes vamos a utilizar el *test de máxima verosimilitud*. Mediante este método podemos estimar los valores de los coeficientes maximizando la probabilidad de obtener los resultados observados. Para poder aplicarlo necesitamos construir una función que se conoce como *función de verosimilitud*. Ésta indica la probabilidad de obtener los datos observados en función de parámetros desconocidos (los β_j-s).
Supongamos otra vez que trabajamos con una sola variable X; por tanto, debemos estimar únicamente los coeficientes β_0 y β_1. Partiendo de una muestra del tipo (x_i, y_i) con $i = 1, 2, ..., n$ individuos, la función de verosimilitud se define como:

$$l(\beta) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i} \tag{1.8}$$

Los valores que maximizan la función son los estimadores de los coeficientes que estabamos buscando. En general, cuando nos referimos a un valor estimado, utilizamos el símbolo ˆ ; en este caso los valores que estimaremos serán $\hat{\beta}_0$ y $\hat{\beta}_1$. Para encontrar estos valores, debemos derivar la función (1.8) e igualarla a cero. Antes de hacer eso, aplicaremos el logaritmo a la función para así obtener el *logaritmo de máxima verosimilitud* (expresión que nos ayudará más adelante a la hora de trabajar con las derivadas):

$$L(\beta) = ln[l(\beta)] = \sum_{i=1}^{n} y_i ln[p(x_i)] + (1 - y_i) ln[1 - p(x_i)] \tag{1.9}$$

Ahora sí, derivando la expresión (1.9) respecto a β_0 y β_1 e igualando los términos a cero, obtenemos las *ecuaciones de máxima verosimilitud*:

$$\sum_{i=1}^{n} y_i - p(x_i) = 0 \tag{1.10}$$

$$\sum_{i=1}^{n} x_i(y_i - p(x_i)) = 0 \tag{1.11}$$

Para resolver estas ecuaciones no lineales necesitamos un método numérico como el método de Newton-Raphson [2].
Una vez resueltas las ecuaciones (1.10) y (1.11), conoceremos los valores $\hat{\beta}_0$ y $\hat{\beta}_1$ que buscabamos, las *estimaciones de máxima verosimilitud*. La estimación de la probabilidad $p(X)$ en función de esos coeficientes es la siguiente:

$$\hat{p} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}} \tag{1.12}$$

Volviendo al caso de regresión múltiple, supongamos que partimos de una muestra de n observaciones independientes del tipo $(\mathbf{x}_i, y_i)$, $i = 1, 2, ...n$, para estimar el vector $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$. Utilizaremos el mismo método que en la regresión simple, pero en lugar de la variable X, ahora utilizaremos el vector $\mathbf{X}$ y en lugar de la probabilidad $p(X)$, el valor de probabilidades $p(\mathbf{X})$.
Estas son las funciones de máxima verosimilitud que obtenemos:

$$\sum_{i=1}^{n} [y_i - p(\mathbf{x}_i)] = 0 \tag{1.13}$$

$$\sum_{i=1}^{n} x_{ij}[y_i - p(\mathbf{x}_i)] = 0, \tag{1.14}$$

con *i=1, ..., n* y *j=1,...,q.*

Resolviendo las ecuaciones conseguimos los $q+1$ coeficientes que buscabamos. Llamaremos $\hat{\boldsymbol{\beta}}$ al vector formado por esos coeficientes.
Vamos a estudiar también los errores estandarizados de los coeficientes que acabamos de estimar. Para ello, siguiendo con esta teoría utilizaremos las segundas derivadas parciales del logaritmo de máxima verosimilitud. Estos valores ayudarán a estimar las varianzas y covarianzas de los coeficientes. Derivando respecto al parámetro j, estas son las expresiones que obtenemos:

$$\frac{\partial L^2(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 p_i(1-p_i) \tag{1.15}$$

$$\frac{\partial L^2(\beta)}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} p_i(1-p_i) \tag{1.16}$$

donde $j, l = 1, ..., q$ y $p_i = p(\mathbf{x}_i)$.

Estos términos forman una matriz de $(p+1) \times (p+1)$ a la que denotaremos $\mathbf{I}(\boldsymbol{\beta})$. Al calcular la inversa de ésta matriz, obtenemos las varianzas y covarianzas de las estimaciones, $\text{Var}(\boldsymbol{\beta})=\mathbf{I}^{-1}(\boldsymbol{\beta})$. La varianza del coeficiente $\hat{\beta}_j$, $\hat{Var}(\hat{\beta}_j)$, es el elemento j-avo de la diagonal de la matriz; y un elemento arbitrario de la no diagonal con posiciones (j, l) con $j, l = 1, ..., p$, nos indica la covarianza de $\hat{\beta}_j$ y $\hat{\beta}_l$, $\hat{Cov}(\hat{\beta}_j, \hat{\beta}_l)$. De esta matriz podemos obtener también el error estándar de la estimación $\hat{\beta}_j$, $\hat{es}(\hat{\beta}_j) = \sqrt{\hat{Var}(\hat{\beta}_j)}$, valor que utilizaremos más adelante.

1.1.2. Significación de las variables

Una vez construido el modelo de regresión logística, nos planteamos si la variable X_j influye en la probabilidad de éxito. Para eso debemos comparar los dos modelos, el que incluye la variable (modA) con el que no (modB).

- Test de máxima verosimilitud

Este método compara los modelos con el llamado test de máxima verosimilitud, utilizando el logaritmo de máxima verosimilitud antes definido. La siguiente expresión se conoce como la *"deviance"* y es la que compara los valores observados con los estimados, basándose en la función de verosimilitud:

$$D = -2ln\left[\frac{Verosimilitud \quad modA}{Verosimilitud \quad modB}\right] \tag{1.17}$$

Sustituyendo la función antes definida en la expresión (1.17) obtenemos esto:

$$D = -2\sum_{i=1}^{n}\left[y_i ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i) ln\left(\frac{1-\hat{p}_i}{1-\hat{y}_i}\right)\right], \tag{1.18}$$

donde $\hat{p}_i = \hat{p}(x_i)$ indica las probabilidades de cada individuo.

Con intención de comprobar la significación de una de las variables independientes, compararemos los valores de la *deviance* al incluir o descartar la variable en la ecuación. Para medir estos valores definiremos el siguiente estadístico G:

$$G = D(modA) - D(modB)$$

El primero, el $modA$, no incluye la variable X_j en el modelo y el $modB$ si que la utiliza. El estadístico G sigue una distribución Chi cuadrado con 1 grado de libertad (hay un único parámetro de diferencia entre los dos modelos), $G \approx \chi^2_1$. Para estudiar la significación de la variable X_j, plantearemos el siguiente contraste de hipótesis:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

donde el p-valor obtenido es p= $P(\chi^2_1 > G)$.

De igual manera, podemos analizar la significación de más de una variable al mismo tiempo. Suponiendo que un modelo tiene r variables y el otro q, siendo $1 < r < q$, veamos cuál es el modelo que mejor se ajusta mediante el siguiente contraste:

$$\begin{cases} H_0 : \beta_{r+1} = ... = \beta_q = 0 \\ H_1 : \exists j \in (r+1, ..., q)/\beta_j \neq 0 \end{cases}$$

El modelo que tiene las q variables es el modB, y el que sólo incluye las r primeras el modA. Este es el estadístico que utilizaremos para resolver el contraste:

$$G = D(modA) - D(modB) = -2ln\left[\frac{Verosimilitud \quad modA}{Verosimilitud \quad modB}\right] \approx \chi^2_{q-r} \tag{1.19}$$

Igualmente, G sigue una distribución Chi cuadrado pero esta vez con $q-r$ grados de libertad. En este caso son $q-r$ los grados de libertad, porque estamos mirando si $q-r$ coeficientes son nulos; en el pasado había un solo β_j y por eso era 1 el grado de libertad. El p-valor será p= $P(\chi^2_{q-r} > G)$.

Antes de comprobar si la variable X_j influye en el modelo, deberíamos asegurarnos de que el vector $\mathbf{X}$ en conjunto es influyente. Para eso realizaremos el siguiente contraste de hípotesis:

$$\begin{cases} H_0 : \beta_1 = ... = \beta_q = 0 \\ H_1 : \exists j \in (1, ..., q)/\beta_j \neq 0 \end{cases}$$

Al modelo que tiene como función la constante β_0 lo llamaremos $modA$ y al que incluye las q variables $modB$. Este es el estadístico que usaremos para resolver el contraste:

$$G = D(modA) - D(modB) = -2ln \left[\frac{Verosimilitud \quad modA}{Verosimilitud \quad modB} \right] \approx \chi^2_q \quad (1.20)$$

G sigue una distribución Chi cuadrado con q grados de libertad y el p-valor que obtenemos es p= $P(\chi^2_q > G)$.

- Test de Wald

Este método también analiza la significación de la variable X_j. Primero, explicaremos el caso de regresión simple para comprender mejor el método, por tanto, debemos estudiar la significación de la variable X.

Como en todos los casos particulares del modelo lineal generalizado, los coeficientes del modelo de regresión logística siguen una distribución normal, es decir, $\hat{\beta}_1 \approx N(\mu, \sigma^2)$. Como se cumplen $\mu = \beta_1$ y σ^2 =Var$(\hat{\beta}_1)$, el estimador β_1 es insesgado y, por tanto, la siguiente fracción sigue una distribución normal estandarizada.

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{es}(\hat{\beta}_1)} \approx N(0, 1) \quad (1.21)$$

Ahora plantearemos el siguiente contraste de hipótesis:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Bajo la hipótesis nula, el estadístico que obtenemos es el siguiente:

$$W_p = \frac{\hat{\beta}_1}{\hat{es}(\hat{\beta}_1)} \quad (1.22)$$

Por tanto, $W_p \approx N(0, 1)$ con un p-valor de p= $2P(Z > |W_p|)$.

Supongamos de nuevo que trabajamos con q variables. En este caso miraremos la significación de la variable X_j, siendo $j \in (1, ..., n)$. Siguiendo los mismos pasos que en el caso univariante, este será el contraste de hipótesis que haremos:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

Bajo la hipótesis nula, el valor del estadístico es:

$$W_p = \frac{\hat{\beta}_j}{\sqrt{\left|I^{-1}(\hat{\beta})\right|_{jj}}} \approx N(0,1) \tag{1.23}$$

siendo p= $2P(Z > |w_p|)$.

Otra manera de analizar la significación de las variables es mediante los intervalos de confianza. Utilizando los valores del test de Wald, estos son los intervalos que obtenemos para el caso univariante:

$$I_{\beta_1}^{1-\alpha} = \left(\hat{\beta}_1 - z_{\alpha/2}\ \hat{es}(\hat{\beta}_1), \hat{\beta}_1 + z_{\alpha/2}\ \hat{es}(\hat{\beta}_1)\right) \tag{1.24}$$

Si contamos con un total de p variables en el modelo, este será el intervalo de confianza para el coeficiente β_j:

$$I_{\beta_j}^{1-\alpha} = \left(\hat{\beta}_j - z_{\alpha/2}\sqrt{\left|I^{-1}(\hat{\beta})\right|_{jj}}, \hat{\beta}_j + z_{\alpha/2}\sqrt{\left|I^{-1}(\hat{\beta})\right|_{jj}}\right)$$

1.1.3. Interpretación de los parámetros

Una vez construido el modelo y visto que las variables son significantes es importante saber interpretar los coeficientes que hemos estimado. Suponiendo que sólo trabajamos con una variable X, vamos a conocer el papel que tiene β_1 en el modelo. Recordemos cuál es el modelo de regresión logística simple:

$$g(X) = \beta_0 + \beta_1 X \tag{1.25}$$

En la regresión logística, el β_1 indica el crecimiento del modelo al aumentar en una unidad el valor de la variable independiente, es decir, $\beta_1 = g(x + 1) - g(x)$. Ahora interpretaremos los coeficientes caso a caso.

- X variable dicotómica

Tal y como hemos comentado anteriormente, el β_1 indica el crecimiento del modelo al aumentar en una unidad el valor de la variable, pero al tratarse de una variable dicotómica, sólo puede tomar dos valores. Supongamos que esos valores son el 0 y el 1. Por tanto, esta es la diferencia del modelo entre dos individuos, uno con $X = 0$ y otro con $X = 1$:

$$g(1) - g(0) = [\beta_0 + \beta_1] - [\beta_0]$$

Para poder interpretar estos resultados tenemos que definir el término *Odds Ratio (OR)*. Llamaremos $p(1)$ a la probabilidad de acierto cuando $X = 1$, y $p(0)$ a la probabilidad de acierto cuando $X = 0$. Es decir,

$$p(1) = P(Y = 1|X = 1) \Rightarrow 1 - p(1) = P(Y = 0|X = 1)$$
$$p(0) = P(Y = 1|X = 0) \Rightarrow 1 - p(0) = P(Y = 0|X = 0)$$

Vamos a definir también el *Odds* de acierto. Si suponemos que la variable X toma el valor 1, $o(Odds \text{ de } X = 1) = \frac{p(1)}{1-p(1)}$; en el caso contrario ($X = 0$), $o(Odds \text{ de } X = 0) = \frac{p(0)}{1-p(0)}$. Al dividir estas dos expresiones, conseguimos el término que buscábamos, el *Odds Ratio*:

$$OR = \frac{p(1)/(1 - p(1))}{p(0)/(1 - p(0))}$$

Sustituyendo los valores en la ecuación (1.25) obtenemos lo siguiente:

$$OR = \frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} / \frac{1}{1+e^{\beta_0+\beta_1}}}{\frac{e^{\beta_0}}{1+e^{\beta_0}} / \frac{1}{1+e^{\beta_0}}} = \frac{e^{2weta_0+\beta_1}}{e^{\beta_0}} = e^{(\beta_0+\beta_1)-\beta_0} = e^{\beta_1}$$

Este valor indica cuánto más frecuente es que ocurra el evento ($Y = 1$) para $X = 1$ que para $X = 0$. En el caso de que $OR = 1$, $\beta_1 = 0$ y por tanto, asumimos que ambas categorías afectan igual a que suceda el acontecimiento $Y = 1$. En el caso de que $OR > 1$, $\beta_1 > 0$, y por tanto, la categoría no de referencia $X = 1$ aumenta la probabilidad de que suceda el evento (Y=1). Por último, si $OR < 1$, la categoría $X = 1$ disminuye la probabilidad de que suceda el evento.

- X variable categórica

En este caso la variable X tendrá más de dos categorías, y por tanto el modelo será un poco diferente. Por ejemplo, si la variable tiene 3 categorías que son $X = 0, 1, 2$, tomando la categoría $X = 0$ como de referencia, este es el modelo que obtenemos:

$$g(X) = logit(X) = \beta_0 + \beta_{11}1\{X = 1\} + \beta_{12}1\{X = 2\} \tag{1.26}$$

De la misma manera definimos los *Odds* de acierto en los tres casos; $p(0)$ con $X = 0$, $p(1)$ con $X = 1$ y $p(2)$ con $X = 2$. Si dividimos el *Odds* de $X = 1$ con el de $X = 0$, obtenemos un *Odds Ratio* de e^{β_1}. Este valor indica cuánto más frecuente es que ocurra el evento (Y=1) para X=1 que para X=0. Por el contrario, si dividimos en *Odds* de $X = 2$ con el de $X = 0$, conseguimos $OR = e^{\beta_2}$. De igual manera, este indica cuánto más frecuente es que ocurra el evento ($Y = 1$) para la categoría $X = 2$ que para la de referencia $X = 0$. Si estos valores son mayores que 1, $e^{\beta_s} > 1$, esa categoría s es la que aumenta la probabilidad de que suceda el evento respecto a $X = 0$; y si son menores que 1, la disminuye. Si $OR = e^{\beta_s} = 1$, $\beta_s = 0$ y por tanto, asumimos que la categoría s y la categoría de referencia afectan de igual modo en la probabilidad de que suceda el evento $Y = 1$, con $s = 1, 2$.

- X variable continua

De manera similar a la anterior, llamaremos $p(x)$ a la probabilidad de acierto cuando $X = x$ y $p(x+1)$ a la probabilidad de acierto cuando $X = x+1$. Es decir,

$$p(x) = P(Y = 1|X = x) \Rightarrow 1 - p(x) = P(Y = 0|X = x)$$
$$p(x+1) = P(Y = 1|X = x+1) \Rightarrow 1 - p(x+1) = P(Y = 0|X = x+1)$$

Definimos también el *Odds* de acierto para este caso. Si suponemos que $X = x$, $o(Odds \text{ de } X = x) = \frac{p(x)}{1-p(x)}$; en el caso de que $X = x+1$, $o(Odds \text{ de } X = x+1) = \frac{p(x+1)}{1-p(x+1)}$. Al dividirlas conseguimos el *Odds Ratio* para variables continuas:

$$OR = \frac{p(x+1)/(1-p(x+1))}{p(x)/(1-p(x))}$$

Sustituyendo y simplificando obtenemos el mismo OR que antes, $OR = e^{\beta_1}$. Este indica cuánto más frecuente es que ocurra el evento (Y=1) al aumentar el valor de X en una unidad, de $X = x$ a $X = x+1$. Si $OR = 1$ ($\beta_1 = 0$), la probabilidad de que suceda el evento es la misma si $X = x$ que si $X = x+1$. Si $OR > 1$, $\beta_1 > 0$; por tanto el aumento de una unidad en X aumenta la probabilidad de que suceda el evento. Y por último, si $OR < 1$, el aumento de una unidad de X disminuye la probabilidad de que suceda el evento (Y=1).

Suponiendo de nuevo que trabajamos con q variables ($\mathbf{X} = (X_1, ..., X_q)$), vamos a interpretar cada uno de los coeficientes β_j. En el caso de que queramos interpretar la variable X_j, debemos entender las $q-1$ variables restantes como constantes, es decir,

$$g(X_j) = (\beta_0 + ... + \beta_{j-1}X_{j-1} + \beta_{j+1}X_{j+1} + ... + \beta_q X_q) + \beta_j X_j$$

Partiendo de esta expresión podemos interpretar el coeficiente β_j igual que hemos hecho antes, dependiendo de si la variable es continua o categórica.

1.2. Bondad de Ajuste y capacidad predictiva del modelo

En este apartado debemos comprobar que las predicciones coinciden con las observaciones. Por un lado, la calibración compara los valores predichos ($\hat{\mathbf{y}} = (\hat{y_1}, \hat{y_2}, ..., \hat{y_n})$) con los observados ($\mathbf{y} = (y_1, y_2, ..., y_n)$) en grupos de individuos, y a medida que la distancia entre ambos disminuye mejora la calibración del modelo. Igualmente, debemos comprobar la discriminación. Diremos que el modelo discrimina bien si el error que obtenemos al comparar cada pareja $(y_i, \hat{y_i})$, con $i = 1, ..., n$, no depende de los valores observados.

1.2.1. Test de Hosmer-Lemeshow

Supongamos que nuestro modelo ajustado consta de q variables, $\mathbf{X} = (X_1, X_2, ..., X_q)$, y que la muestra se obtiene a partir de n individuos. Vamos a empezar por definir los términos que utilizaremos a lo largo de la sección.

J: número de valores distintos que puede observar $\mathbf{X}$.

m_j: número de individuos que cumplen $\mathbf{X} = x_j$ con $j = 1, ..., J \rightarrow \sum_{j=1}^{J} m_j = n$.

y_j: de los m_j individuos, aquellos que cumplen $Y = 1$.

Supongamos que se cumple $J = n$. Crearemos n columnas, cada una con la probabilidad estimada de cada individuo, ordenadas de menor a mayor. Estas columnas se dividen en g grupos siguiendo estos dos criterios:

-los percentiles de las probabilidades estimadas, y

-los valores fijos de las probabilidades estimadas.

Un valor frecuentemente utilizado para g es 10.

El estadístico que crearon Hosmer y Lemeshow es el siguiente:

$$\hat{c} = \sum_{k=1}^{g} \frac{(O_k - n'_k \bar{p_k})^2}{n'_k \bar{p_k}(1 - \bar{p_k})} \tag{1.27}$$

donde n'_k indica el número de combinaciones con las variables del grupo k, $O_k = \sum_{j=1}^{n'_k} y_j$ es el número de aciertos ($Y = 1$) del grupo n_k, y $\bar{p_k} = \sum_{j=1}^{n'_k} \frac{m_j \hat{p_j}}{n'_k}$ la media de la probabilidad estimada.

Después de hacer muchas simulaciones, Hosmer y Lemeshow demostraron que si $J = n$ y el modelo de regresión logística es el correcto, el estadístico $\hat{c}$ se aproxima con una distribución Chi cuadrado de $g - 2$ grados de libertad. Es decir, $\hat{c} \approx \chi^2_{g-2}$ y en este caso el p-valor será p= $P(\chi^2_{g-2} > \hat{c})$. En la práctica, se asume que esta misma distribución aproxima bien el estadístico cuando $J \approx n$.

1.2.2. Tablas de clasificación

Con intención de resumir la información que obtenemos del modelo ajustado, vamos a crear las llamadas tablas de clasificación. En estas tablas, se resume mucha de la información que conseguimos a partir de un modelo. Para poder crear las tablas, necesitamos hablar de los puntos de corte.

En este caso, estamos trabajando con una variable respuesta dicotómica y, por tanto, necesitamos concretar cuándo vamos a tomar las probabilidades estimadas como de éxito ($\hat{Y} = 1$) o de fracaso ($\hat{Y} = 0$). Esa selección la haremos a partir del punto de corte c. Si $\hat{p}(\mathbf{X}) \geq c$ tomaremos $\hat{Y} = 1$ y, en cambio, si $\hat{p}(\mathbf{X}) < c$, asumiremos que $\hat{Y} = 0$.

Una vez elegido el punto de corte podemos empezar a construir la tabla que compara los valores observados (Y) con los estimados ($\hat{Y}$). En total tenemos n individuos y llamaremos $a+b$ al número de fracasos observados y $c+d$ a los aciertos Tabla 1.1. Apreciese que los valores observados son fijos, mientras que los estimados varían con el punto de corte c. A partir de esta tabla podemos conocer la siguiente información:

	Estimados ($\hat{Y}$)		
Observados (Y)	0	1	
0	a	b	a+b
1	c	d	c+d
	a+c	b+d	n

Tabla 1.1: Tabla de clasificación, sensibilidad y especificidad

La proporción obtenida a partir de los éxitos que se aciertan (d), es decir, los que cumplen $Y = \hat{Y} = 1$, se obtiene de la expresión $\frac{d}{c+d} \times 100$ y se conoce como la $\frac{d}{c+d}$ **sensibilidad**.

La que se consigue a partir de los fracasos que se aciertan (a), es decir, los que cumplen $Y = \hat{Y} = 0$, se obtiene de la expresión $\frac{a}{a+b} \times 100$ y se conoce como la $\frac{a}{a+b}$ **especificidad**.

La **proporción global de aciertos** se basa en la siguiente expresión $\frac{a+d}{n} \times 100$.

La sensibilidad y la especificidad se pueden interpretar de esta otra manera.

Probabilidad de acierto:

- Sensibilidad $\Rightarrow Se(c) = \mathrm{P}(\hat{\pi}(\mathbf{X}) \geq c | Y = 1)$
  Sabiendo que sucede el acierto ($Y = 1$), indica la probabilidad de que las probabilidades estimadas para el éxito sean mayores o iguales que c.

- Especificidad $\Rightarrow Sp(c) = \mathrm{P}(\hat{\pi}(\mathbf{X}) < c | Y = 0)$
  Expresa la probabilidad de que las probabilidades estimadas sean menores al punto de corte c, en los individuos que cumplen $Y = 0$.

Probabilidad de error:

- 1 - sensibilidad $\Rightarrow \; 1 \; - \; Se(c) = \mathrm{P}(\hat{\pi}(\mathbf{X}) < c | Y = 1)$
  Conociendo que sucede el acierto ($Y = 1$), indica la probabilidad de que las probabilidades estimadas para el éxito sean menores que c.

- 1 - especificidad $\Rightarrow$ $1 - Sp(c) = \mathrm{P}(\hat{\pi}(\mathbf{X}) \geq c | Y = 0)$
  Revela la probabilidad de que las probabilidades estimadas sean mayores o iguales al punto de corte c, en los individuos que cumplen $Y = 0$.

1.2.3. La curva ROC

La curva ROC (Receiver Operating Characteristic), es una curva que relaciona la sensibilidad con la (1-especificidad) para un conjunto entero de puntos de corte. En este caso, las probabilidades estimadas toman valores en el intervalo $(0, 1)$ por lo que, $c \in (0, 1)$.

$$ROC(.) = \{((1 - Sp(c)), Se(c)), c \in (0, 1)\} \tag{1.28}$$

La función ROC está definida en $[0, 1] \times [0, 1]$, es monótona y ascendente. Podemos representarla también de la siguiente forma:

$$ROC(.) = \{t, ROC(t), t \in (0, 1)\} \tag{1.29}$$

La Figura 1.1 muestra dos curvas ROC diferentes:

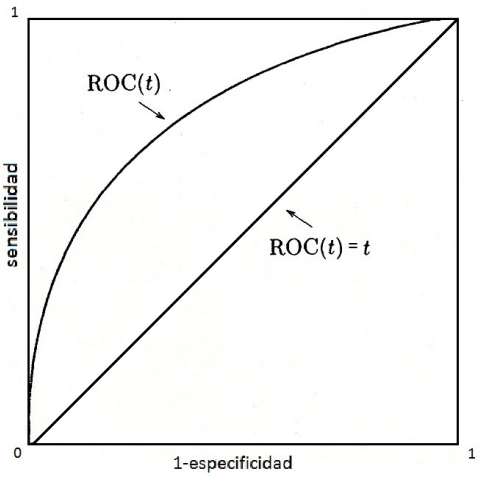

Figura 1.1: Dos ejemplos de curvas ROC

La curva ROC$(t) = t$, es la referente a una variable que no tiene capacidad discriminante, es decir, se considera que no es capaz de predecir el éxito o el fracaso de la variable respuesta. Se compara esta situación con el tirar una moneda al aire y predecir si caerá del lado de la cara o de la cruz. En cambio la curva perfecta, es la que cumple $(1 - Sp(c)) = 0$ y $Se = 1$ para algún valor de c. Esta curva en su totalidad es muy próxima a la parte superior

e izquierda del cuadrado de la imagen y cumple $(1 - Sp(c)) > Se(c)$ para todos los valores de c.

Normalmente una curva cualquiera ROC(t) se encuentra entre las dos curvas que acabamos de mencionar. En el caso de que queramos comparar dos curvas gráficamente, consideraremos mejor la curva que más se aproxime a la perfecta.

El AUC

El AUC (Area Under the ROC Curve) es un parámetro que mide la capacidad discriminativa de un modelo, y como su propio nombre indica, se calcula a partir de la curva ROC de la siguiente manera:

$$AUC = \int_0^1 \mathrm{ROC(r)}\,\mathrm{d}r \tag{1.30}$$

La variable o el modelo ídilico será aquel con una curva perfecta, es decir, un AUC de 1.0 (valor máximo que puede tomar); mientras que aquel que no aporte información (la curva ROC(t)=t) tendrá un AUC de 0.5. Esta última variable no será capaz de predecir el éxito o el fracaso del individuo.

Para entender mejor el concepto, vamos a comparar el área bajo la curva de dos variables X_A y X_B. Suponiendo que X_A es mejor, vamos a comparar las dos variables graficamente:

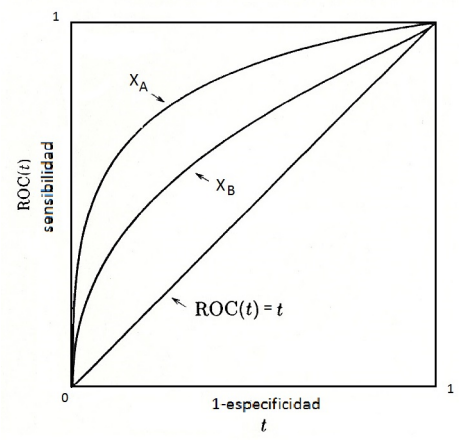

Figura 1.2: Curva ROC para las variables X_A y X_B

La curva que más se aproxime a la curva perfecta, en este caso la de la variable X_A, será la correspondiente al modelo o variable que tenga mejor capacidad discriminante y su AUC será mayor que el de la variable X_B.

Matemáticamente, estos son los resultados que obtenemos para una variable X_A mejor que otra variable X_B.

$$ROC_{X_A}(t) \geq ROC_{X_B}(t) \qquad \forall t \in (0,1) \tag{1.31}$$

Igualmente el AUC obtendrá mejores resultados, un valor más cercano a 1:

$$AUC_{X_B} < AUC_{X_A} \leq 1 \tag{1.32}$$

En la práctica, estudiaremos si hay diferencias estadísticamente significativas entre dos AUCs mediante el test de DeLong [9]. Este es el contraste de hipótesis que se plantea en el test:

$$\begin{cases} H_0 : \text{AUC}_{X_A} = \text{AUC}_{X_B} \\ H_1 : \text{AUC}_{X_A} \neq \text{AUC}_{X_B} \end{cases}$$

1.3. Sistema de puntuación para la creación de un score

Al trabajar con modelos predictivos en el terreno de la bioestadística, parece interesante conocer cómo influye cada variable que forma el modelo (factores de riesgo, enfermedades, etc.) en la variable respuesta Y. El objetivo es crear un score que prediga la variable respuesta en función del modelo multivariante que se ha obtenido. Utilizaremos la metodología propuesta por Sullivan et al [3], que resumimos a continuación.

(i) Estimar los parámetros del modelo multivariante

Consideramos el modelo $g(p(\mathbf{X})) = \beta_0 + \beta_1 X_1 + ... + \beta_q X_q$ de regresión logística, donde $p(\mathbf{X}) = E(Y|X)$ indica la esperanza de que suceda Y a partir de esas variables $\mathbf{X}$. El modelo $g(p(\mathbf{X}))$ relaciona la esperanza de que suceda el evento Y como combinación de los factores de riesgo $X_1, ..., X_q$, y $\beta_0, \beta_1, ..., \beta_q$ son los coeficientes estimados $\hat{\beta}_j$ basados en el test de máxima verosimilitud.

(ii) Dividir los factores de riesgo en categorías y determinar cada valor de referencia

Nos limitaremos a explicar como trabajar con las variables categóricas, ya que todas las variables que formarán nuestro modelo serán de este tipo. Normalmente, se dividen los resultados en dos categorías: valores normales ($X_j = 0$) y valores de riesgo ($X_j = 1$), aunque también pueden incluirse variables con más de dos categorías; estos niveles se diseñan siguiendo parámetros médicos.

Tras haber dividido los factores de riesgo en niveles, debemos indicar cuál es el valor de referencia de cada uno de ellos. Denotaremos W_{js} al valor de referencia de la categoría s de la variable j (X_j), para $j = 1, ..., q$ y $s = 1, ..., c_j$, siendo c_j el número de categorías de la variable j (X_j). Por ejemplo, si la variable j indica la presencia o ausencia de una enfermedad, estos serán los valores de referencia: 0 = ausente (W_{j1}), 1 = presente (W_{j2}).

(iii) Determinar la categoría base de cada factor de riesgo

A continuación debemos determinar cuál es la categoría base o de referencia de cada variable j, a la que llamaremos W_{jREF}, con $j = 1, ..., q$. Dicho nivel corresponderá a la categoría de menor gravedad, normalmente $X_{js} = 0$. En el sistema de puntuación, a las categorías de referencia se les asignarán 0 puntos, mientras que los niveles con valores de riesgo tendrán una puntuación positiva. Si se diera el caso de que una categoría no de referencia mostrara mejores valores que W_{jREF}, se le asignaría una puntuación negativa en el score.

(iv) Establecer la distancia a la que está cada categoría respecto de la categoría base en unidades de regresión

Necesitamos conocer la distancia a la que se encuentran los valores de referencia de cada categoría del valor de referencia de la categoría base. En términos de la regresión logística esta es la distancia de la que hablamos:

$$\beta_j(W_{js} - W_{jREF}), \quad \text{con } j = 1, ..., q \text{ y } s = 1, ..., c_j.$$

(v) Fijar el multiplicador fijo o constante B

Ahora debemos definir la constante para el sistema de puntuación; valor que coincide con el número de unidades del modelo de regresión logística que corresponde a un punto del sistema.

(vi) Establecer el número de puntos que se le asignan a cada categoría de los factores de riesgo

Una vez conocemos la unidad de nuestro sistema de puntuación, podemos dar puntos a las diferentes categorías de los distintos factores de riesgo:

$$\text{Puntos}_{js} = \beta_j(W_{js} - W_{jREF})/B,$$

para cada variable j y categoría s (de las c_j categorías que tiene j). Se aprecia que la puntuación para las categorías de referencia es de 0 puntos.

(vii) Determinar el riesgo asociado a la puntuación total

Es necesario conocer la puntuación máxima que puede tomar un individuo, para así clasificar las puntuaciones en distintos niveles de riesgo. Dependiendo de la puntuación asignada a cada nivel, el score tendrá una puntuación máxima u otra. Para poder dar el último paso en la creación del score, necesitamos estimar el riesgo o la probabilidad de que el paciente fallezca. Ese riesgo estimado p varia dependiendo del modelo de regresión que hayamos utilizado, en el caso de la regresión logística, esta es la probabilidad de mortalidad estimada:

$$p(\mathbf{X}) = \frac{e^{\sum_{j=0}^{q} \beta_j X_j}}{1 + e^{\sum_{j=0}^{q} \beta_j X_j}} \tag{1.33}$$

La idea del sistema de puntuación es aproximar la contribución de cada factor de riesgo en la estimación del riesgo $p(\mathbf{X})$. Se sabe que el producto de la puntuación total del score y la constante B, aproxima $\sum_{j=1}^{q} \beta_j X_j$, pero debemos tener en cuenta el intercepto. Por tanto,

$$\sum_{j=0}^{q} \beta_j X_j \approx \beta_0 + B(\text{puntuación total})$$

Mediante la creación del score podemos aproximar la probabilidad de que un evento suceda, de que un paciente fallezca, por ejemplo, de una manera muy fácil y rápida. Por eso son muy utilizados en el ámbito clínico.

1.4. Validación del score

A la hora de desarrollar un score a partir de un modelo predictivo, es importante acreditar la validez del mismo. Para eso, se debe probar que el score también es buen predictor al aplicarlo a nuevos datos [4].

Hay diversos métodos para llevar a cabo la validación de un score. En el caso de contar con una segunda muestra para hacer la validación, se habla de validación externa; mientras que si partimos de la misma base para crear y validar el score, se le llama validación interna. Hay diversas técnicas para llevar a cabo la validación interna, como por ejemplo la *cross-validation*, *bootstrap validation* o la *split-sample validation* [4].
Nosotros trabajaremos con el método *split-sample validation*. Esta técnica se basa en dividir la muestra total en dos submuestras, la de derivación y la de validación. La submuestra de derivación se utiliza pata desarrollar el modelo, mientras que la de validación se usa para validarlo. La muestra se divide de forma aleatoria, normalmente tomando el 50 % para una y el restante 50 % para la otra, o el 60 % para una y el 40 % para la otra.

El uso de muestras de validación ayuda a valorar la capacidad predictiva del score en la segunda muestra. Compararemos el área bajo la curva ROC

del score en cada muestra, y mediante el test de Delong [9] decidiremos si hay diferencias estadísticamente significativas entre los AUCs en ambas muestras.

Capítulo 2

Aplicación a la Neumología

2.1. Objetivo

La finalidad de este trabajo es crear un score a partir de un modelo multivariante de regresión logística que ayude a predecir si un paciente hospitalizado por neumonía fallece o no durante su ingreso. La idea es facilitar a los médicos un sistema mediante el cual haciendo un número reducido de pruebas a los pacientes, puedan determinar la gravedad de los mismos. Para ello, se han recogido datos de 1.388 pacientes ingresados en el Hospital Universitario de Cruces en el periodo de 2005 a 2011.

Por otro lado, también nos interesa comparar los resultados obtenidos a partir de nuestro modelo, con otros dos scores que usan los neumólogos a día de hoy: el CURB-65 y el FINE. Tal y como veremos más adelante, las variables que forman el CURB-65 se recogen facilmente en una consulta rutinaria, mientras que algunos de los factores de riesgo que forman el FINE, requieren de distintas pruebas, como por ejemplo de un análisis de sangre. Por eso, crearemos dos scores distintos. El primer score al que llamaremos *Quick-Decision Score* (QDS), ayudará al especialista a tomar una decisión en apenas 15 minutos, lo que dura una consulta rutinaria; mientras que el otro, el *Analitic-Based Score* (ABS), lo formarán variables que requieren de diversos análisis de sangre. Cada uno de los scores tiene un aspecto positivo: mientras que el segundo es más completo, el primer score permite clasificar al paciente con mayor rapidez.

2.2. Descripción de la base de datos

La base de datos la forman una larga lista de variables que aportan la siguiente información acerca de cada individuo: datos demográficos como la *edad* o el *sexo*; antecedentes médicos tales como las enfermedades previas (la *diabetes*, por ejemplo) o el *tabaquismo*; datos clínicos y exploratorios del

paciente como la *temperatura*, el *estado mental*, la *frecuencia respiratoria* y la *cardiaca*, la *presión arterial*, etc.; datos analíticos obtenidos tras una prueba de sangre como por ejemplo la *PCR* (marcador de inflamación), la PaO_2 (mide la presión del oxígeno en sangre), la prueba del *BUN* (prueba de urea), la del *pH*, etc.; alteraciones radiológicas que valoran la extensión de la neumonía como la *afectación multilobar* o la existencia de *derrame pleural* (presencia de líquido en la membrana que rodea ambos pulmones, en la pleura); tratamientos empleados (como los *antibióticos* o la *ventilación mecánica*) y los scores CURB-65 y FINE (la base de datos recoge tanto los niveles de riesgo en los que cada score divide al paciente, como las variables que los forman). Más adelante explicaremos al detalle estos dos scores para luego así poder compararlos con los que creemos nosotros.

Como ya hemos dicho, nos interesa ayudar a los expertos a clasificiar a los enfermos en distintos grupos de riesgo en el menor tiempo posible. Por tanto, de todas las variables disponibles en el estudio se han considerado aquellas clínicamente relevantes para estimar la mortalidad. Para decidir cuáles son esas variables, contamos con la opinión experta de uno de los médicos del equipo de neumólogos del Hospital Universitario de Cruces.
A continuación, explicaremos con más detalle los factores que los médicos tienen en consideración al explorar a los pacientes y el tipo de pruebas que se les realizan, indicando también cuales son los valores de riesgo de cada variable. Dividiremos las variables en dos tipos: por un lado, aquellas relativas a la situación previa del paciente (enfermedades previas, edad, etc.) que aumentan el riesgo de que padezca neumonía y de que esta sea más grave; y, por otro lado, criterios que indiquen la gravedad de la neumonía:

- Antecedences personales

- *Edad*. De los datos demográficos recogidos tomaremos este para nuestro modelo porque, claramente, la edad del individuo influye en la aparición y desarrollo de la neumonía. Una de las causas originarias de la enfermedad son las bacterias o los virus, y estas enfermedades son más propensas entre las personas mayores.

- *Asilo*. Esta variable indica si el paciente vive en una residencia o no. Se ha visto que los pacientes que viven en un asilo tienen infecciones por bacterias más agresivas.

- *Demencia*. Indica si el paciente presenta deterioro mental previo a la neumonía o no.

- *Enfermedad pulmonar crónica*. Estas enfermedades aumentan el riesgo de infecciones pulmonares como la neumonía. Por ello, estudiaremos esta variable dicotómica que indica si el enfermo tiene algún tipo de enfermedad pulmonar crónica o no.

- *Insuficiencia cardíaca*. La variable muestra si el paciente ha sufrido o sufre insuficiencia cardíaca, o no. Estudiaremos también la variable que indica si el paciente ha sufrido un *infarto* o no.

- *Neoplasia activa o en el último año*. La variable determina si el paciente padece cáncer (o lo ha sufrido en el último año) o no.

- *Enfermedad cerebrovascular*. Define si el paciente ha sufrido un accidente cerebrovascular anterior (infarto cerebral, hemorragia, etc.) o no.

- *Diabetes*. Esta variable indica si el enfermo es diabético o no, es decir, si tiene la glucemia (glucosa en sangre) elevada.

- *Insuficiencia renal crónica*. Muestra si el paciente tiene enfermedad renal previa o no. Otra variable que revela información sobre el estado renal del paciente es la *alteración renal*. Esta variable indica si el paciente ha sufrido alguna complicación renal o no.

- *Hepatopatía*. Señala si el enfermo padece una enfermedad previa del hígado o no.

- Criterios de gravedad

- *Alteración del estado mental*. Esta variable muestra si el paciente presenta alteración mental o no. El especialista deberá valorar si el paciente se encuentra desorientado con respecto a las personas, el tiempo o el lugar en el que se encuentra; respondiendo afirmativa o negativamente a esta cuestión.

- *Temperatura*. En la primera exploración el médico tomará la temperatura del paciente.

- *Frecuencia respiratoria*. Esta variable recoge el número de veces que el enfermo respira por minuto. Es una prueba rápida de realizar y por eso la estudiaremos para nuestro modelo predictivo. Un buen resultado sería de 12-16 rpm (respiraciones por minuto), mientras que uno superior se considera taquipnea; de 16 a 24 respiraciones taquipnea leve, de 24 a 30 moderada, y grave si es superior a 30.

- *Frecuencia cardíaca*. Esta variable recoge el número de latidos por minuto (lpm). Consideraremos como factor de riesgo una frecuencia mayor de 120 lpm.

- *Presión arterial sistólica*. Corresponde al valor máximo de la tensión arterial cuando el corazón se contrae. La prueba de la tensión arterial mide la presión existente en los vasos sanguíneos y se obtiene mediante una prueba exploratoria. Una presión baja supone un fallo del sistema circulatorio, y mayor gravedad para el paciente. La prueba se recoge en milímetros de mercurio (mm Hg) y se entiende como anómalo un resultado inferior a 95 mm Hg.

- *Presión arterial diastólica.* Aunque normalmente se utiliza la prueba anterior, también es interesante conocer la tensión arterial cuando el corazón se relaja. Se conoce como presión arterial diastólica al valor mínimo de la tensión arterial entre latidos. Los resultados de la prueba son también inmediatos y por eso tan útiles. Entenderemos como normal una presión arterial diastólica mayor que 60 mm Hg.

- *Presión de oxígeno en sangre (PaO_2).* Este dato obtenido tras una prueba analítica del plasma, mide la presión del oxígeno en sangre y se recoge en milímetros de mercurio. Se considerará a un paciente como de riesgo cuando la presión de oxígeno en sangre sea menor que 60 mm Hg.

- *Saturación de oxígeno en sangre.* Es una prueba que ofrece resultados similares a la anterior. Su obtención es más fácil y rápida por lo que se utiliza mucho. Se mide en tanto por ciento y se tiene como de riesgo un resultado inferior a 90 %.

- *Creatinina.* Los problemas renales son un factor en la gravedad de la neumonía, y por eso, estudiaremos esta prueba que mide la función renal. Esta prueba se realiza a partir de un análisis de sangre y se considera que un resultado superior o igual a 1.8 mg/dl es un mal resultado. La base de datos recoge también la variable *BUN (Blood Urea Nitrogen)*, que como su nombre indica corresponde al nitrógeno ureico en sangre. Tanto el CURB-65 como el FINE utilizan esta última en sus scores.

- *pH.* Indica el equilibrio ácido-base sanguíneo. Se considera normal un resultado entre 7.35 y 7.45. Cuando la cifra es menor, el paciente se encuentra en acidosis (dato de gravedad). La acidosis puede ser tanto metabólica como respiratoria.

- *Sodio.* Es un ión disuelto en plasma. Sus cifras normalmes son entre 135 y 145 mEq/l (mili equivalente por litro).

- *Glucosa.* Mediante una prueba de sangre se determina la glucosa (azúcar) del paciente. En una persona sana, se considera un nivel elevado más de 110 mg/dl (miligramo por decilitro).

- *Hematocrito.* Refleja la cantidad de hemoglobina (proteína transportadora de oxígeno) en sangre. Se mide en tanto por ciento. Se consideran valores normales 42-52 % en hombres y 37-47 % en mujeres. Generalmente se entiende como de riesgo un valor menor al 30 %.

- *Derrame pleural.* Tras una radiografía, podremos determinar si hay líquido en la pleura (membrana que rodea los pulmones) o no, prueba que determina si el paciente tiene un derrame pleural o no.

- *Afectación multilobar.* Mediante una radiografía, se observará si el paciente presenta problemas en más de un lóbulo o no. Debe tenerse en cuenta que cada sistema respiratorio consta de 5 lóbulos, 3 en el pulmón derecho y 2 en

el izquierdo. Se considera que un paciente tiene neumonía cuando tiene al menos uno de esos cinco lóbulos afectado.

- *Número de lóbulos afectados.* Esta variable indicará el número de lóbulos dañados de uno a cinco. La situación del paciente será peor a mayor número de lóbulos afectados.

La base de datos recoge también la puntuación de los dos scores con los que vamos a comparar nuestro modelo. Para facilitar la comprensión del estudio que han hecho al crear estos sistemas de puntuación, vamos a explicarlos con un poco más de detalle.

2.2.1. CURB-65

Este marcador toma como variable respuesta la mortalidad a los 30 días desde el ingreso, constando unicamente de cuatro variables, las cuales le dan nombre: **CURB** (**C**onfusion, **U**rea, **R**espiratory rate, **B**lood pressure) [5]. Normalmente se incluye también la variable que indica la edad del paciente, aunque no se incluye como continua sino como dicotómica. El punto de corte que utilizan para categorizarla es el de **65** años y por eso el score se llama CURB-65.
Este marcador es uno de los más utilizados por varias razones. Por una parte, los resultados de las cinco variables que lo forman son rápidos de conseguir; por otro lado, los enfermos son divididos en tres categorías y dependiendo en cúal se encuentre, el médico tomará una decisión u otra. Como en la mayoría de casos clínicos, las cinco variables que forman este score están categorizadas. Todas ellas se dividen en dos grupos: si $x_j = 1$ entenderemos el factor como de riesgo, y si $x_j = 0$ no, con $j = 1, ..., 5$. Los puntos de corte utilizados son los que se entienden medicamente como de riesgo. Entenderemos como valores de riesgo los siguientes:

- Signos de confusión mental (puntuación menor o igual a 8 en el Score del Test Mental, o desorientación en la persona, tiempo o lugar).
- BUN superior a 20 mg/dl (prueba de urea).
- Frecuencia respiratoria mayor o igual que 30 pulsaciones por minuto.
- Tensión arterial sistólica menor que 90 mm Hg, o tensión arterial diastólica menor o igual que 60 mm Hg.
- Edad mayor o igual a 65 años.

Teniendo en cuenta el número de factores de riesgo que tiene el paciente se clasifica en uno de estos tres grupos:

- **Baja mortalidad** (0 ó 1): Estos pacientes tienen una puntuación de 0 ó 1 en el score y no se consideran graves. Probablemente, el paciente es adecuado para el tratamiento en domicilio.

- **Mortalidad intermedia** (2): Este grupo lo forman los pacientes con 2 factores de riesgo de los 5 que forman el score. Se realizará un tratamiento supervisado por el hospital; el especialista deberá decidir si el paciente necesita ser hospitalizado o si puede considerarse paciente externo con revisiones periódicas.
- **Alta mortalidad** (3 o más): Los enfermos que muestren 3 o más signos de los que se consideran de riesgo se consideran graves. Por tanto se hospitalizará al paciente con un diagnóstico de neumonía grave. Si la puntuación es de 4 o 5, se evaluará la necesidad de que sea valorado por la UCI (Unidad de Cuidados Intensivos).

Nuestra base de datos recoge las puntuaciones del CURB-65 y también las distintas pruebas para dar puntuación al score.

2.2.2. FINE

Este marcador también toma como variable respuesta la mortalidad tras un mes desde el ingreso [6]. El primer paso para conocer la gravedad del paciente es la realización de un análisis inicial por parte del doctor. El especialista deberá responder sí o no a las siguientes tres cuestiones:

- ¿Tiene el individuo más de 50 años?
- ¿Tiene el enfermo un historial reciente de alguna de las siguientes condiciones: enfermedad neoplásica, fallo cardíaco congestivo, enfermedad cerebrovascular, enfermedad renal o enfermedades hepáticas?
- ¿Sufre el paciente alguna de las siguientes anomalías: alteración del estado mental, frecuencia cardíaca $\geq$125 lpm, frecuencia respiratoria $\geq$30 rpm, tensión arterial sistólica $<$90 mm Hg, o temperatura $<$35°o $\geq$40°?

Si la respuesta a las tres preguntas es negativa el médico asignará al paciente al grupo de riesgo I. En cambio, si alguna de las respuestas es afirmativa, el doctor deberá realizar más pruebas. A partir de los resultados obtenidos, se calculará la puntuación correspondiente al paciente de acuerdo a la siguiente tabla:

Tabla 2.1: Escala de FINE para la NAC (Neumonía adquirida por la comunidad)

Continúa en la siguiente página.

Edad (hombres)	edad (años)
Edad (mujeres)	edad (años) - 10
Asilo, residencia	+ 10
Neoplasia	+ 30
Enfermedad hepática	+ 20
Fallo cardíaco congestivo	+ 10
Enfermedad cerebrovascular	+ 10
Enfermedad renal	+ 10
Alteración del estado mental	+ 20
Frecuencia respiratoria $\geq$ 30 rpm	+ 20
Tensión arterial sistólica $<$ 90 mm Hg	+ 20
Temperatura $<$ 34 o $>$ 40	+ 15
Frecuencia cardíaca $\geq$ 125 lpm	+ 10
pH $<$ 7.35	+ 30
BUN $\geq$ 30 mg/dl	+ 20
Sodio $<$ 130 mg/dl	+ 20
Glucosa $>$ 250 mg/dl	+ 10
Hematocrito $<$ 30 %	+ 10
PO_2 $<$ 60 mm Hg	+ 10
Derrame pleural	+ 10

Una vez obtenida la puntuación final, se clasificará al paciente entre los grupos de riesgo II a V. Por tanto, de acuerdo con los resultados de las pruebas el paciente estará en una de estas categorías:

- **Grupo de riesgo I**. Lo forman los enfermos menores a 50 años, que no sufran ninguna de las comorbilidades antedichas ni obtengan resultados anómalos en sus pruebas. El paciente no necesita ser ingresado y será enviado a su domicilio.
- **Grupo de riesgo II**. Se incluirán en este grupo los pacientes que hayan obtenido una puntuación inferior a 70 puntos. El estado de estos no requiere hospitalización.
- **Grupo de riesgo III**. Esta categoría estará formada por los que obtengan entre 71 y 90 puntos. En estos enfermos se deberá valorar el ingreso hospitalario.
- **Grupo de riesgo IV**. Este conjunto lo forman los que tengan de 91 a 130 puntos, debiendo hospitalizar a estos pacientes.
- **Grupo de riesgo V**. De obtener una puntuación superior a 130 puntos, el paciente será clasificado como de riesgo muy alto. Tal es su gravedad, que deberá ser valorado por la UCI.

2.3. Análisis univariante

En esta sección se presenta la aplicación de la metodología explicada en el Capítulo 1 a los datos presentados en la Sección 2.2. Se ha utilizado el sofware estadístico R [7] para el análisis de datos. Concretamente, se han utilizado las funciones `roc` y `roc.test` del paquete `pROC` [8] para la estimación y comparación de AUCs respectivamente.

Ahora que conocemos el objetivo del trabajo y las variables que formarán parte del estudio, podemos empezar a trabajar para crear los dos modelos predictivos. Antes de nada, debemos recordar que dividiremos aleatoriamente los n individuos en dos muestras distintas (una del 60 % (derivación) y la otra del 40 % (validación), aproximadamente) creando los scores basándonos en la parte de derivación, reservando la otra para la posterior validación del modelo.

Antes de crear el modelo predictivo de mortalidad en la muestra de derivación, debemos comprobar que la proporción de individuos que fallece es la misma en ambas muestras (derivación y validación). En la siguiente tabla mostramos los porcentajes de los individuos que fallecen en ambas muestras:

	Mortalidad		Total
	No	Si	
Derivación	96.42 %	3.58 %	867
Validación	95.39 %	4.61 %	521

p-valor=0.340

Mediante un test de Chi cuadrado obtenemos un p-valor superior a 0.05, por lo que no rechazamos la hipótesis nula y por tanto, admitimos que la proporción de individuos que fallece es la misma en ambas muestras. Por tanto, de ahora en adelante, trabajaremos con la muestra de derivación, que recoge los resultados de 867 individuos.

Una vez hemos seleccionado las variables clínicamente relevantes, nos aseguraremos de que también lo sean matemáticamente; es decir, comprobaremos la significación de cada una respecto a la variable respuesta mortalidad. Los resultados de algunas de las pruebas médicas con las que contamos se recogen de forma continua, pero en base a criterio y recomendación clínica las incluiremos de forma categórica. Por lo general, las dividiremos en dos categorías: $X_j = 0$ cuando los valores sean normales, no de riesgo, y $X_j = 1$ cuando los resultados se consideren anómalos. Los puntos de corte que determinan qué valores son de riesgo y cuáles no, los fijaremos a partir de la recomendación de un neumólogo del Hospital Universitario de Cruces.

La tabla descriptiva que aparece a continuación (Tabla 2.2), muestra cuántos individuos fallecen y cuántos no, en cada categoría de cada variable X_j en la muestra de derivación.

Tabla 2.2: Tabla descriptiva de cada variable y número de valores perdidos en la muestra de derivación.

		Mortalidad	
	N.A. (%)	No	Si
Edad	0		
<65 años		373 (97.64 %)	9 (2.36 %)
≥65 años		463 (95.46 %)	22 (4.54 %)
Asilo	0		
No		819 (96.58 %)	29 (3.42 %)
Si		17 (89.47 %)	2 (10.53 %)
Demencia	0		
No		813 (96.67 %)	28 (3.33 %)
Si		23 (88.46 %)	3 (11.54 %)
Enfermedad pulmonar crónica	0		
No		636 (97.55 %)	16 (2.45 %)
Si		200 (93.02 %)	15 (6.97 %)
Insuficiencia cardíaca	2 (0.24 %)		
No		758 (96.8 %)	25 (3.2 %)
Si		76 (92.68 %)	6 (7.32 %)
Neoplasia activa	2 (0.24 %)		
No		800 (96.39 %)	30 (3.61 %)
Si		34 (97.14 %)	1 (2.86 %)
Enfermedad cerebrovascular	0		
No		780 (96.42 %)	29 (3.58 %)
Si		56 (96.55 %)	2 (3.45 %)
Infarto	1 (0.12 %)		
No		797 (96.72 %)	27 (3.28 %)
Si		38 (90.48 %)	4 (9.52 %)
Diabetes	3 (0.35 %)		
No		669 (97.52 %)	17 (2.48 %)
Si		164 (92.13 %)	14 (7.87 %)
Insuficiencia renal crónica	2 (0.24 %)		
No		782 (96.66 %)	27 (3.34 %)
Si		30 (93.75 %)	2 (6.25 %)
Alteración renal	0		
No		777 (96.64 %)	27 (3.36 %)
Si		59 (93.65 %)	4 (6.35 %)
Hepatopatía	0		
No		807 (96.99 %)	25 (3.01 %)
Si		29 (82.86 %)	6 (17.14 %)
Alteración del estado mental	1 (0.12 %)		
No		770 (97.6 %)	19 (2.4 %)
Si		65 (84.42 %)	12 (15.58 %)

Continúa en la siguiente página.

Tabla 2.2: Tabla descriptiva de cada variable y número de valores perdidos en la muestra de derivación.

		Mortalidad	
	N.A. (%)	No	Si
Temperatura	0		
$< 39.5°$		785 (96.44 %)	29 (3.56 %)
$\geq 39.5°$		51 (96.23 %)	2 (3.77 %)
Frecuencia respiratoria	1 (0.12 %)		
$<$ 24 rpm		552 (99.46 %)	3 (0.54 %)
$\geq$ 24 rpm		283 (91 %)	28 (9 %)
Frecuencia cardíaca	0		
$\leq$ 120 lpm		746 (97.13 %)	22 (2.87 %)
$>$ 120 lpm		90 (90.9 %)	9 (9.1 %)
Presión arterial sistólica	0		
$\geq$ 95 mm Hg		760 (97.56 %)	19 (2.44 %)
$<$ 95 mm Hg		76 (86.36 %)	12 (13.64 %)
Presión arterial diastólica	0		
$>$ 60 mm Hg		594 (97.86 %)	13 (2.14 %)
$\leq$ 60 mm Hg		242 (93.08 %)	18 (6.92 %)
Presión arterial de O_2 (PaO_2)	180 (20.76 %)		
$\geq$ 60 mm Hg		413 (97.41 %)	11 (2.59 %)
$<$ 60 mm Hg		244 (92.78 %)	19 (7.22 %)
Saturación de O_2 en sangre	4 (0.46 %)		
$\geq$ 90 %		696 (97.89 %)	15 (2.11 %)
$<$ 90 %		137 (90.13 %)	15 (9.87 %)
Creatinina	0		
$<$ 1.8 mg/dl		756 (97.17 %)	22 (2.83 %)
$\geq$ 1.8 mg/dl		80 (89.89 %)	9 (10.11 %)
pH	180 (20.76 %)		
$\geq$ 7.35		642 (97.27 %)	18 (2.73 %)
<7.35		15 (55.56 %)	12 (44.44 %)
Sodio	0		
$\in$ [135-145) mEq/l		571 (97.27 %)	16 (2.73 %)
$\notin$ [135-145) mEq/l		265 (94.64 %)	15 (5.36 %)
Glucosa	0		
$<$ 175 mg/dl		691 (97.19 %)	20 (2.81 %)
$\geq$ 175 mg/dl		145 (92.95 %)	11 (7.05 %)
Hematocrito	1 (0.12 %)		
$\geq$ 30 %		809 (96.42 %)	30 (3.58 %)
$<$ 30 %		26 (96.3 %)	1 (3.7 %)
Derrame pleural	0		
No		770 (96.61 %)	27 (3.39 %)
Si		66 (94.29 %)	4 (5.71 %)
Afectación multilobar	0		
No		668 (97.66 %)	16 (2.34 %)
Si		168 (91.8 %)	15 (8.2 %)

Continúa en la siguiente página.

Tabla 2.2: Tabla descriptiva de cada variable y número de valores perdidos en la muestra de derivación.

		Mortalidad	
	N.A. (%)	No	Si
Número de lóbulos afectados	0		
1		668 (97.66 %)	16 (2.34 %)
2		140 (97.22 %)	4 (2.78 %)
3 o 4		28 (71.79 %)	11 (28.21 %)
5		0 (0 %)	0 (0 %)

Si nos fijamos en la variable que indica el *número de lóbulos afectados*, vemos que ningún paciente tiene todos los lóbulos dañados; por otro lado, el por centage de enfermos que fallece en la primera categoría, es muy similar al de la segunda.

Una vez que conocemos las variables con las que haremos el estudio vamos a verificar si son significativas respecto a la variable respuesta Y de forma univariante. La Tabla 2.3 muestra los p-valores que obtenemos para las 28 variables en un modelo univariante con una significación de 0.05 en base al test de máxima verosimilitud. Aparecen en negrita los p-valores de las variables que hemos seleccionado para el estudio; las que cumplen p <0.05, o las que tienen relevancia médica con un p-valor menor a 0.2.

Tabla 2.3: Modelos univariantes: p-valor de cada variable en la muestra de derivación

Variable X_j	p-valor	variable X_j	p-valor
Edad	**0.080**	Asilo	0.175
Demencia	**0.073**	Enfermedad pulmonar crónica	**0.004**
Insuficiencia cardíaca	**0.087**	Neoplasia activa	0.807
Enfermedad cerebrovascular	0.956	Infarto	0.073
Diabetes	**0.002**	Insuficiencia renal crónica	0.961
Alteración renal	0.261	Hepatopatía	**<0.001**
Alteración del estado mental	**<0.001**	Temperatura	0.937
Frecuencia respiratoria	**<0.001**	Frecuencia cardíaca	**0.006**
Presión arterial sistólica	**<0.001**	Presión arterial diastólica	**<0.001**
Presión arterial de O_2	**0.005**	Saturación del O_2 en sangre	**<0.001**
Creatinina	**0.003**	pH	**<0.001**
Sodio	**0.058**	Glucosa	**0.018**
Hematocrito	0.972	Derrame pleural	0.350
Afectación multilobar	**<0.001**	Número de lóbulos afectados	**<0.001**

La Tabla 2.4 resume la información que obtenemos de esas variables; muestra los coeficientes $\hat{\beta}$ estimados, los *odds ratio*, los intervalos de confianza de los odds ratio al 95 %, el AUC de cada variable y el p-valor que obtenemos

mediante el test de *razón de verosimilitud* en el modelo univariante. En todos los casos se ha considerado la categoría de referencia aquella categoría de menor riesgo.

Tabla 2.4: Coeficiente β de cada modelo univariante, el *Odds Ratio* (OR), el intervalo de confianza del OR, el AUC y el p-valor de cada variable en la muestra de derivación.

	β	OR (IC OR)	AUC	p-valor
Edad			0.578	0.080
$<$ 65 años	-	-		
$\geq$ 65 años	0.677	1.969 (0.896, 4.325)		
Demencia			0.535	0.073
No	-	-		
Si	1.332	3.787 (1.074, 13.36)		
Enfermedad pulmonar crónica			0.622	0.004
No	-	-		
Si	1.09	2.981 (1.448, 6.137)		
Insuficiencia cardíaca			0.551	0.087
No	-	-		
Si	0.873	2.394 (0.952, 6.017)		
Infarto			0.542	0.073
No	-	-		
Si	1.134	3.107 (1.035, 9.328)		
Diabetes			0.627	0.002
No	-	-		
Si	1.212	3.359 (1.623, 6.956)		
Hepatopatía			0.579	$<$ 0.001
No	-	-		
Si	1.899	6.678 (2.545, 17.528)		
Alteración del estado mental			0.655	$<$ 0.001
No	-	-		
Si	2.013	7.481 (3.479, 16.091)		
Frecuencia respiratoria			0.782	$<$ 0.001
$<$24 rpm	-	-		
$\geq$24 rpm	2.902	18.205 (5.487, 60.402)		
Frecuencia cardíaca			0.591	0.006
$\leq$ 120 lpm	-	-		
$>$ 120 lpm	1.221	3.391 (1.515, 7.589)		
Presión arterial sistólica			0.603	$<$ 0.001
$\geq$ 95 mm Hg	-	-		
$<$ 95 mm Hg	1.859	6.415 (2.713, 15.181)		
Presión arterial diastólica			0.656	$<$ 0.001
$>$ 60 mm Hg	-	-		
$\leq$ 60 mm Hg	1.223	3.399 (1.164, 7.042)		
Presión arterial de O_2 (PaO_2)			0.631	0.005
$\geq$ 60 mm Hg	-	-		
$<$ 60 mm Hg	1.073	2.924 (1.368, 6.248)		

Continúa en la siguiente página.

Tabla 2.4: Coeficiente β de cada modelo univariante, el *Odds Ratio* (OR), el intervalo de confianza del OR, el AUC y el p-valor de cada variable en la muestra de derivación.

	β	OR (IC OR)	AUC	p-valor
Saturación de O_2 en sangre			0.668	< 0.001
$\geq 90\,\%$	-	-		
$< 90\,\%$	1.625	5.080 (2.427, 10.635)		
BUN			0.746	< 0.001
$\leq$ 20 mg/dl	-	-		
$>$ 20 mg/dl	0.052	1.053 (1.033, 1.073)		
Creatinina			0.597	0.003
<1.8 mg/dl	-	-		
≥1.8 mg/dl	1.352	3.866 (1.721, 8.682)		
pH			0.689	< 0.001
≥7.35	-	-		
<7.35	3.351	28.533 (11.694, 69.623)		
Sodio			0.583	0.109
$\in$ [135-145) mEq/l	-	-		
$\notin$ [135-145) mEq/l	0.703	2.020 (0.984, 4.147)		
Glucosa			0.591	0.008
$<$ 175 mg/dl	-	-		
$\geq$ 175 mg/dl	0.964	2.621 (1.229, 5.589)		
Afectación multilobar			0.642	< 0.001
No	-	-		
Si	1.316	3.728 (1.804, 7.629)		
Número de lóbulos afectados			0.668	< 0.001
1	-	-		
2	0.1422	1.153 (0.380, 3.499)		
3 o 4	2.791	16.304 (6.929, 38.365)		

Una vez hemos resumido la información que aporta cada variable de manera univariante, vamos a decidir cuáles tomamos para que formen parte de nuestro modelo predictor completo (con análisis de sangre).
Debemos tener en cuenta que las variables que incluyamos en nuestro modelo multivariante, deben ser independientes entre ellas, es decir, cada variable debe aportar distinta información sobre el estado del paciente. Es por eso por lo que debemos descartar algunas de ellas.

Es el caso de las variables *afectación multilobar* y *número de lóbulos afectados*. La primera indica si el paciente tiene un solo lóbulo afectado (afectación multilobar = No), o si tiene 2 o más dañados (afectación multilobar = Si); mientras que la segunda indica el número de lóbulos dañados de 1 a 5. La segunda variable recoge la información de la primera especificando cuántos lóbulos tiene dañados en el caso de que haya afectación multilobar. Por eso, descartaremos la variable *afectación multilobar* y trabajaremos únicamente con la que indica el *número de lóbulos afectados*.
La prueba del *BUN* y la de la *creatinina* recogen resultados similares; ambas

pruebas miden la urea en sangre. En este caso utilizaremos la *creatinina* puesto que es la utilizada por los médicos hoy en día.
La *presión arterial sistólica* y la *presión arterial diastólica* son pruebas que estan bastante relacionadas, aunque se utiliza más la primera. Nosotros incluiremos la *sistólica*, la misma que incluyen los scores CURB-65 y FINE, pero con otro punto de corte.
Hay dos pruebas que sirven para determinar la presión del oxígeno en la sangre: la *presión de oxígeno en sangre* (PaO_2) y la *saturación de oxígeno en sangre*. La prueba del PaO_2 tiene más de un 20 % de valores perdidos, y la de la saturación es más facil y rápida de obtener, por lo que utilizaremos la segunda.
La variable que indica si el paciente ha sufrido algún *infarto* obtiene un p-valor de 0.073 con el test de máxima verosimilitud, es decir, con una significación α de 0.05 la rechazamos para nuestro modelo multivariante. Aunque sucede lo mismo con la variable *insuficiencia cardíaca*, incluiremos esta porque es más global que la anterior, y por eso parece más interesante trabajar con ella.
Las variables *glucosa* y *diabetes* están relacionadas porque los pacientes con antecedentes de diabetes tienen glucemia elevada, aunque en ocasiones, pacientes sin diabetes también la pueden tener. Hemos incluido como variable el diagnóstico previo o no de diabetes, porque esta enfermedad predispone a enfermedades como la neumonía.

Por tanto, seleccionaremos las siguientes variables que aportan información sobre los antecedentes personales del paciente: *edad* mayor o igual a 65 años, *demencia*, si ha sufrido *enfermedades pulmonares crónicas* o *insuficiencia cardíaca*, y si tiene *diabetes* o alguna *hepatopatía*.
Respecto a las que indican la gravedad del paciente, incluiremos las siguientes: signos de *alteración del estado mental*, resultados anómalos en las pruebas de *frecuencia respiratoria* ($\geq$24 rpm), *frecuencia cardíaca* (>120 lpm), *presión arterial sistólica* (<95 mm Hg), *saturación en sangre* (<90 %), *creatinina* ($\geq$ 1.8 mg/dl), *pH* (<7.35), *sodio* (<135 o $\geq$145 mEq/l) y de *glucosa* ($\geq$175 mg/dl), y tras una placa, la variable que indica el *número de lóbulos afectados*.

2.4. Desarrollo del *Quick-Decision Score* (QDS)

Como hemos explicado en los objetivos, vamos a crear un score que ayude a determinar la gravedad del paciente de neumonía en el menor tiempo posible. Para saber si un paciente tiene neumonía, es necesarion realizar una radiografía que determine si tiene algún lóbulo afectado, ya que únicamente puede concluirse que el paciente padece neumonía tras comprobar que al menos uno de los lóbulos está dañado. Por tanto, siempre contaremos con

los resultados necesarios para dar valor al factor de riesgo que indica el número de lóbulos afectados (de 1 a 5).
Realizaremos un primer modelo que recoja los factores de riesgo cuyos valores puedan conocerse en una consulta rutinaria de unos 15 minutos, es decir, los que no necesiten de un análisis de sangre.

2.4.1. Desarrollo del modelo *Quick-Decision*

Tras haber seleccionado las variables que formarán parte de este primer estudio, queremos construir el modelo que las relaciona con la variable respuesta mortalidad. Para eso, debemos estimar los coeficientes β_j para cada variable X_j, en este caso con $j = 1, ..., 12$. En la siguiente tabla mostramos las estimaciones $\hat{\beta}_j$ y los p-valores que obtenemos con el test de Wald, al igual que los odds ratio al 95 % e intervalos de confianza del OR de cada variable.

Tabla 2.5: Modelo multivariante ajustado en la muestra de derivación con las variables seleccionadas en el análisis univariante (sin incluir las variables que requieren de pruebas analíticas)

	β	OR	IC OR	p-valor
Intercept	−7.54	-		<0.001
Edad				
<65	-	-	-	-
≥65	0.905	2.471	(0.736-8.294)	0.143
Demencia				
No	-	-	-	-
Si	1.038	2.825	(0.513-15.567)	0.233
Enfermedad pulmonar				
No	-	-	-	-
Si	1.188	3.280	(1.222-8.802)	0.018
Insufuciencia cardíaca				
No	-	-	-	-
Si	0.450	1.568	(0.497-4.949)	0.443
Diabetes				
No	-	-	-	-
Si	1.126	3.082	(1.186-8.001)	0.021
Hepatopatía				
No	-	-	-	-
Si	1.531	4.622	(1.126-18.970)	0.034
Alteración del estado mental				
No	-	-	-	-
Si	0.901	2.479	(0.816-7.529)	0.109

Continúa en la siguiente página.

Tabla 2.5: Modelo multivariante ajustado en la muestra de derivación con las variables seleccionadas en el análisis univariante (sin incluir las variables que requieren de pruebas analíticas)

	β	OR	IC OR	p-valor
Frecuencia respiratoria				
<24	-	-	-	-
≥24	2.250	9.485	(2.448-36.754)	0.001
Frecuencia cardíaca				
≤120	-	-	-	-
>120	0.300	1.349	(0.417-4.360)	0.617
Tensión arterial sistólica				
≥95	-	-	-	-
<95	1.306	3.692	(1.281-10.641)	0.016
Saturación del O_2 en sangre				
≥90	-	-	-	-
<90	0.595	1.813	(0.700-4.695)	0.220
Número de lóbulos afectados				
1	-	-	-	-
2	−0.492	0.611	(0.135-2.766)	0.523
3 o 4	2.734	15.388	(3.962-59.770)	< 0.001

La intención es crear un modelo predictor que esté formado por un conjunto de variables, todas ellas significativas. Como vemos en la Tabla 2.5, no todos los p-valores son menores a 0.05, por tanto, no todas son significativas. Empezaremos por eliminar del modelo la variable referente a la *frecuencia cardíaca* por tener el mayor p-valor.
Cada vez que eliminamos una variable del modelo comparamos el modelo viejo y el nuevo mediante el test de razón de verosimilitud como hemos explicamos en la Sección 1.1.2.
Seguiremos descartanto variables hasta que consigamos que todos los factores de riesgo que forman el modelo sean significativos. Así obtenemos el modelo predictor definitivo (Tabla 2.6):

Tabla 2.6: Modelo multivariante formado por las variables significativas ajustado en la muestra de derivación con las variables seleccionadas en el análisis univariante (sin incluir las variables que requieren de pruebas analíticas)

	β	OR	IC OR	p-valor
Intercept	−6.635	-	-	<0.001

Continúa en la siguiente página.

Tabla 2.6: Modelo multivariante formado por las variables significativas ajustado en la muestra de derivación con las variables seleccionadas en el análisis univariante (sin incluir las variables que requieren de pruebas analíticas)

	β	OR	IC OR	p-valor
Enfermedad pulmonar				
No	-	-	-	-
Si	1.318	3.737	(1.484-9.411)	0.005
Diabetes				
No	-	-	-	-
Si	1.222	3.394	(1.353-8.512)	0.009
Alteración del estado mental				
No	-	-	-	-
Si	1.207	3.342	(1.211-9.224)	0.020
Frecuencia respiratoria				
<24	-	-	-	-
≥ 24	2.422	11.268	(3.216-39.481)	<0.001
Tensión arterial sistólica				
≥ 95	-	-	-	-
<95	1.479	4.388	(1.609-11.968)	0.004
Número de lóbulos afectados				
1	-	-	-	-
2	-0.559	0.572	(0.138-2.370)	0.441
3 o 4	2.723	15.221	(4.780-28.466)	< 0.001

2.4.2. Creación del score QDS

Una vez contamos con el modelo predictivo, podemos empezar a dar puntuación a los distintos factores de riesgo para así crear el score. Como hemos comentado anteriormente, la categoría número de lóbulos afectados 2, no es estadísticamente significativa con respecto a la categoría de referencia (1 sólo lóbulo dañado) pero la variable en su conjunto si es significativa (p-valor < 0.001). Por lo tanto, a esa segunda categoría se le asignarán 0 puntos en el score QDS por no ser significativa.

Como hemos visto en el apartado 1.4. de la teoría, una vez construido el modelo de regresión, debemos organizar los factores de riesgo en categorías y determinar cuál es el valor de referencia de cada una. En este caso, todos los factores están divididos en niveles (categorías) y ya tienen el valor de referencia estipulado (No/ Si, 0/ 1, 1/ 2/ 3.5($=\frac{3+4}{2}$)). Tomaremos como categoría base, aquella con menor riesgo; es decir, la categoría a la que se le asignan 0 puntos en el sistema de puntuación; por ejemplo, *No* en el caso de *enfermedad pulmonar*, *0 (*<24*)* en el caso de *frecuencia respiratoria* o *1* en el caso de *número de lóbulos afectados*.

Factor de riesgo	Categorías	W_{js}	β_j	$\beta_j(W_{js} - W_{jREF})$
Enfermedad pulmonar			1.318	
	No	0=W_{1REF}		0
	Si	1=W_{11}		1.318
Diabetes			1.222	
	No	0=W_{2REF}		0
	Si	1=W_{21}		1.222
Alteración del estado mental			1.207	
	No	0=W_{3REF}		0
	Si	1=W_{31}		1.207
Frecuencia respiratoria			2.422	
	<24	0=W_{4REF}		0
	≥24	1=W_{41}		2.422
Presión arterial sistólica			1.479	
	≥95	0=W_{5REF}		0
	<95	1=W_{51}		1.479
Número de lóbulos afectados			2.723	
	1	1=W_{6REF}		0
	2	2=W_{61}		0
	3 o 4	3.5=W_{62}		6.808

Es necesario conocer el número de unidades de regresión que corresponde a un punto del sistema de puntuación, constante a la que llamaremos B. Este valor coincide con el coeficiente β_j menor del modelo predictivo. En este caso, tomamos $B = \beta_3 = 1{,}207$ referente al factor de riesgo que indica si hay alteración en el estado mental. Ahora sí, procederemos a dar puntuación a las distintas categorías.

Factor de riesgo	Categorías	$\beta_j(W_{js} - W_{jREF})$	Puntos$_{js}$= $\beta_j(W_{js} - W_{jREF})/B$
Enfermedad pulmonar			
	No	0	0
	Si	1.318	1
Diabetes			
	No	0	0
	Si	1.222	1
Alteración del estado mental			
	No	0	0
	Si	1.207	1
Frecuencia respiratoria			
	<24	0	0
	≥24	2.422	2

Continúa en la siguiente página.

Factor de riesgo	Categorías	$\beta_j(W_{js} - W_{jREF})$	Puntos$_{js}$= $\beta_j(W_{js} - W_{jREF})/B$
Presión arterial sistólica			
	≥95	0	0
	<95	1.479	1
Número de lóbulos afectados			
	1	0	0
	2	0	0
	3 o 4	6.808	6

Ahora, vamos a determinar el riesgo de cada categoría asociado al número total de puntos del score. Sabemos que la probabilidad estimada para que el individuo fallezca ($Y = 1$), se estima de la siguiente manera:

$$p(\mathbf{X}) = \frac{e^{\sum_{j=0}^{q} \beta_j X_j}}{1 + e^{\sum_{j=0}^{q} \beta_j X_j}}$$

El número total de puntos multiplicado por la constante (B=1.207) aproxima el valor de $\sum_{j=1}^{q} \beta_j X_j$, pero necesitamos $\sum_{j=0}^{q} \beta_j X_j$ y por tanto debemos añadir la estimación del intercepto $\beta_0 = -6.635$. Para este modelo, esta es la aproximación del término:

$$\sum_{j=0}^{q} \beta_j X_j \approx -6.635 + B(\text{total de puntos})$$

Siguiendo esta aproximación del término, calculamos la estimación del riesgo dependiendo de la puntuación obtenida en el sistema. Apreciese que cada individuo obtendrá una puntuación de entre 0 y 12 puntos.

Puntuación	Probabilidad de mortalidad estimada	Puntuación	Probabilidad de mortalidad estimada
0	0.0013	7	0.8598
1	0.0044	8	0.9535
2	0.0145	9	0.9856
3	0.0468	10	0.9956
4	0.1410	11	0.9987
5	0.3543	12	0.9996
6	0.6472		

En el caso de que a un enfermo no se le hayan realizado las pruebas pertinentes para obtener los resultados de los seis factores de riesgo que forman QDS, no podremos calcular la puntuación referente a ese paciente. Por ello, este score se ha creado a partir de los pacientes que no tienen valores perdidos (en esas seis variables).

2.4.3. Validación del QDS

Ahora que conocemos la puntuación que le asigna el score QDS a cada paciente, vamos a comprobar si es buen predictor de la mortalidad al aplicarlo a nuevos datos: es decir, vamos a validar el *quick-decision score* (QDS). Para eso, compararemos el AUC del QDS al aplicarlo en la muestra de derivación y en la de validación, y analizaremos si hay diferencias estadísticamente significativas entre ambas muestras.

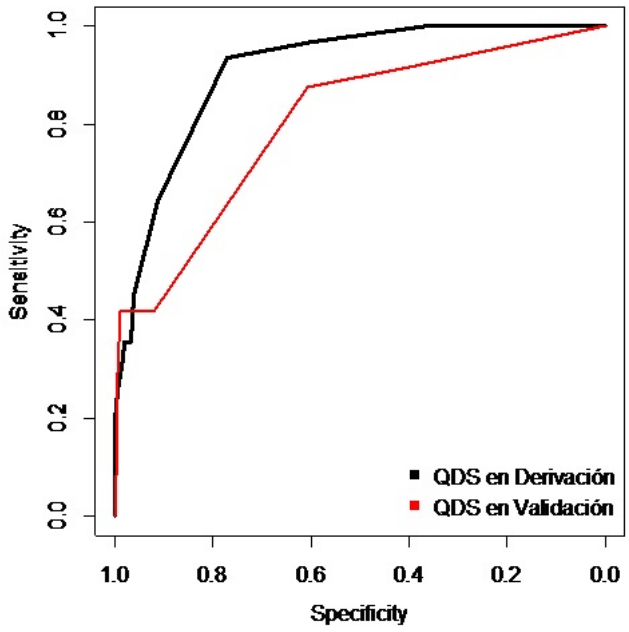

Figura 2.1: Curvas ROC del score QDS en derivación y en validación

La Figura 2.1 muestra la curva ROC del QDS en la muestra de derivación y en la de validación. Se aprecia que aunque el score obtiene buenos resultados en ambas muestras, son mejores en la muestra de derivación (AUC 0.910) que en la de validación (AUC 0.803), y al comparar ambas curvas, observamos que si hay diferencias estadísticamente significativas entre ambas muestras (p-valor de 0.044). De todas maneras, el socre QDS obtiene un AUC superior a 0.8 en la muestra de validación, y por tanto, podemos decir que tiene buena capacidad discriminante.

2.4.4. Comparativa con la literatura

Como hemos dicho en los objetivos una vez creado el score, vamos a compararlo con otros dos que utilizan los expertos: el CURB-65 y el FINE. La base

de datos no recoge la puntuación que le asigna cada score a cada paciente, sino el grupo de riesgo en el que se encuentra (de 1 a 3 en el CURB-65 y de 1 a 5 en el FINE). Por eso, crearemos dos nuevas variables que recojan la puntuación de cada score (de 0 a 5 en el CURB-65 y de 0 a >130 en el FINE).

Vamos a comparar los 3 marcadores en la muestra completa por dos razones: el número de individuos que forman la muestra total es mayor al de las otras (derivación y validación); y, ninguno de los scores ha sido creado en dicha muestra. La siguiente figura representa las curvas ROC de los tres scores en la muestra completa.

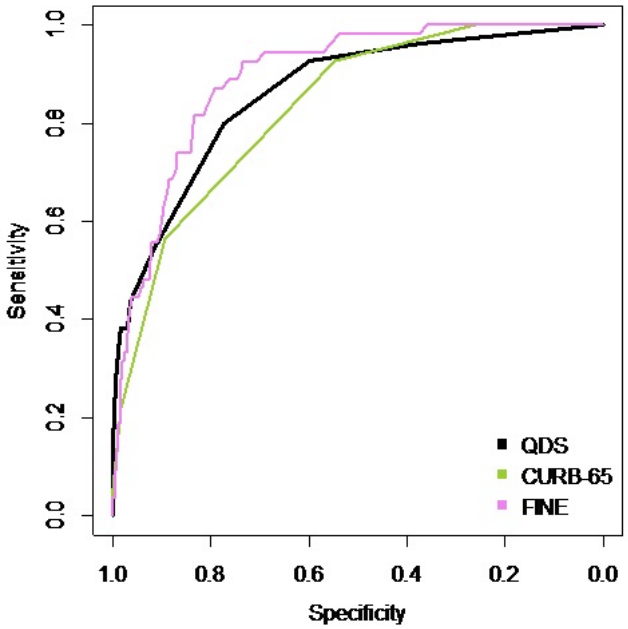

Figura 2.2: Curvas ROC de los tres scores en la muestra completa

Estas son las áreas bajo las curvas ROC (los AUCs) de cada uno de los scores en la muestra completa: 0.866 la del QDS, 0.8319 la del CURB-65 y 0.8936 la del FINE. Si comparamos los resultados de dos en dos, los p-valores que obtenemos son mayores a 0.05 (0.379 al comparar el AUC de QDS con el de CURB-65 y 0.315 al comparar el de QDS con el del FINE); por tanto, no hay diferencias estadísticamente significativas al aplicar cualquiera de los tres scores en la muestra completa.

Una vez hemos creado el QDS y comparado con los otros dos scores, vamos a dividir las distintas puntuaciones en diferentes grupos de riesgo. Dividiremos

a los pacientes en 4 categorías dependiendo de la puntuación que obtengan en el score QDS. Los puntos de corte para categorizar el score se han calculado utilizando la función `catpredi` del paquete `CatPredi` [10].

La siguiente tabla muestra cuántos individuos fallecen y cuántos no en cada categoría del *quick-decision score* en la muestra total.

		No fallece	Fallece
Leve (I)	(0-1 puntos)	796 (99.5 %)	4 (0.5 %)
Moderado (II)	(2 puntos)	231 (97.06 %)	7 (2.94 %)
Grave (III)	(3-6 puntos)	278 (92.36 %)	23 (7.64 %)
Muy grave (IV)	(7-12 puntos)	22 (51.16 %)	21 (48.84 %)

Basándonos en la puntuación que obtiene cada paciente, podemos distinguirlos en estas cuatro categorías:

- **Grupo I**: Los pacientes de esta categoría tienen buen pronóstico y pueden tratarse en domicilio.
- **Grupo II**: Si el individuo obtiene 2 puntos en el sistema de puntuación debe valorarse una estancia corta en el hospital o tratamiento en domicilio con supervisión médica.
- **Grupo III**: Los pacientes de esta categoría necesitan ingreso hospitalario.
- **Grupo IV**: En estos pacientes se debe valorar el ingreso en UCI.

Ahora que contamos con la variable que divide a los pacientes en categorías dependiendo de su gravedad, vamos a calcular el área bajo la curva ROC de esta variable categórica. Al dibujar la curva en la muestra completa, conseguimos un AUC de 0.8553 y por tanto, hemos conseguido crear un score con buena capacidad discriminativa.

2.5. Desarrollo del *Analitic-Based Score* (ABS)

Una vez conseguido el primer modelo predictivo, vamos a crear otro un poco más completo que incluya pruebas que obtenemos a partir de varios análisis de sangre. Por un lado, necesitamos un análisis de sangre rutinario para conocer los resultados de la *creatinina* y del *sodio* (un análisis de sangre venoso). Por otro lado, se le realizará una gasometría al paciente, otro tipo de análisis de sangre que requiere analizar la sangre de la arteria, no de cualquier vaso sanguíneo; mediante esta prueba obtenemos los valores del *pH* y de la *presión arterial del oxígeno en sangre* (PaO_2). Como hemos dicho, la gasometría es una prueba un poco más difícil y dolorosa de realizar, por

eso la variable *pH* tiene bastantes valores perdidos. Será el experto el que decida, en cada caso, si es necesario llevar a cabo esta prueba o no, aunque de presentar el paciente resultados anómalos en la prueba de la *saturación del oxígeno en sangre*, lo habitual es realizar la gasometría.

2.5.1. Desarrollo del modelo *Analitic-Based*

Para crear este modelo más completo, vamos tomar las 17 variables que hemos seleccionado para el estudio. En la siguiente tabla mostramos las estimaciones de los coeficientes β_j-s, los p-valores que obtenemos con el test wald para cada categoría y los odds ratio al 95 % e intervalos de confiaza del OR de cada variable.

Tabla 2.10: Modelo multivariante ajustado en la muestra de derivación con las variables seleccionadas en el análisis univariante (incluyendo las variables que requieren de pruebas analíticas)

	β	OR	IC OR	p-valor
Intercept	−7.759	-	-	<0.001
Edad				
<65	-	-	-	-
≥65	0.628	1.874	(0.504-6.696)	0.348
Demencia				
No	-	-	-	-
Si	1.346	3.843	(0.599-24.641)	0.156
Enfermedad pulmonar				
No	-	-	-	-
Si	1.175	3.237	(1.074-9.756)	0.037
Insuficiencia cardíaca				
No	-	-	-	-
Si	0.676	1.966	(0.580-6.662)	0.278
Diabetes				
No	-	-	-	-
Si	1.257	3.514	(1.209-10.210)	0.021
Hepatopatía				
No	-	-	-	-
Si	1.8362	6.272	(1.134-34.704)	0.035
Alteración del estado mental				
No	-	-	-	-
Si	0.305	1.356	(0.367-5.015)	0.648
Frecuencia respiratoria				
<24	-	-	-	-
≥24	2.210	9.119	(2.911-41.344)	0.004

Continúa en la siguiente página.

Tabla 2.10: Modelo multivariante ajustado en la muestra de derivación con las variables seleccionadas en el análisis univariante (incluyendo las variables que requieren de pruebas analíticas)

	β	OR	IC OR	p-valor
Frecuencia cardíaca				
≥120	-	-	-	-
>120	−0.535	0.589	(0.1343-2.555)	0.477
Presión arterial sistólica				
≥95	-	-	-	-
<95	1.576	4.837	(1.551-15.084)	0.006
Saturación del O_2 en sangre				
≥90	-	-	-	-
<90	0.210	1.234	(0.423-3.595)	0.700
Creatinina				
<1.8	-	-	-	-
≥1.8	−0.085	0.918	(0.256-3.271)	0.895
pH				
≥7.35	-	-	-	-
<7.35	2.699	14.865	(3.595-61.468)	< 0.001
Sodio				
≥135 o <145	-	-	-	-
<135 o ≥145	1.089	2.971	(1.071-8.239)	0.036
Número de lóbulos afectados				
1	-	-	-	-
2	−1.040	0.354	(0.066-1.889)	0.224
3 o 4	2.418	11.223	(2.335-53.950)	0.002

Tal y como hemos hecho con el modelo anterior, vamos a ir descartando las variables que no sean significativas con el test de máxima verosimilitud (p-valor mayor a 0.05). La siguiente tabla muestra las 7 variables que forman el modelo definitivo, las mismas que en el anterior (exceptuando la *alteración del estado mental*) además de los resultados de las pruebas del *sodio* y del *pH*.

Tabla 2.11: Modelo multivariante formado por las variables significativas ajustado en la muestra de derivación con las variables seleccionadas en el análisis univariante (incluyendo las variables que requieren de pruebas analíticas)

	β	OR	IC OR	p-valor
Intercept	−6.762	-		<0.001

Continúa en la siguiente página.

Tabla 2.11: Modelo multivariante formado por las variables significativas ajustado en la muestra de derivación con las variables seleccionadas en el análisis univariante (incluyendo las variables que requieren de pruebas analíticas)

	β	OR	IC OR	p-valor
Enfermedad pulmonar				
No	-	-	-	-
Si	1.224	3.400	(1.256-9.202)	0.016
Diabetes				
No	-	-	-	-
Si	1.296	3.653	(1.333-10.013)	0.012
Frecuencia respiratoria				
<24	-	-	-	-
≥ 24	2.070	7.927	(2.155-29.163)	0.002
Presión arterial sistólica				
≥ 95	-	-	-	-
<95	1.659	5.254	(1.789-15.424)	0.003
pH				
≥ 7.35	-	-	-	-
<7.35	2.917	18.484	(5.145-66.407)	< 0.001
Sodio				
≥ 135 o <145	-	-	-	-
<135 o ≥ 145	0.966	2.628	(1.011-6.829)	0.047
Número de lóbulos afectados				
1	-	-	-	-
2	-1.281	0.278	(0.056-1.379)	0.117
3 o 4	2.379	10.791	(3.140-37.088)	< 0.001

2.5.2. Creación del score ABS

Partiendo de este modelo vamos a seguir los mismos pasos que en el apartado anterior para crear un segundo score. En primer lugar, vamos a indicar cuáles son los niveles de cada factor de riesgo, cuál es la categoría de referencia, y la distancia de cada una de las categorías con respecto a la base.

Factor de riesgo	Categorías	W_{js}	β_j	$\beta_j(W_{js} - W_{jREF})$
Enfermedad pulmonar			1.224	
	No	$0=W_{1REF}$		0
	Si	$1=W_{11}$		1.224
Diabetes			1.296	
	No	$0=W_{2REF}$		0
	Si	$1=W_{21}$		1.296

Continúa en la siguiente página.

Factor de riesgo	Categorías	W_{js}	β_j	$\beta_j(W_{js} - W_{jREF})$
Frecuencia respiratoria			2.070	
	<24	0=W_{3REF}		0
	≥24	1=W_{31}		2.070
Presión arterial sistólica			1.659	
	≥95	0=W_{4REF}		0
	<95	1=W_{41}		1.659
pH			2.197	
	≥ 7.35	0=W_{5REF}		0
	< 7.35	1=W_{51}		2.197
Sodio			0.966	
	∈[135-145)	0=W_{6REF}		0
	∉[135-145)	1=W_{61}		0.966
Número de lóbulos afectados			2.379	
	1	1=W_{7REF}		0
	2	2=W_{71}		0
	3 o 4	3.5=W_{72}		5.948

Sabemos que el número de unidades de regresión que corresponde a un punto del sistema de puntuación, coincide con el coeficiente β_j menor del modelo. En este caso, la constante B es la referente al sodio, por tener menor coeficiente; por tanto, $B = \beta_6$=0.966. Una vez conocemos la unidad, podemos dar puntos a los distintos niveles de cada factor de riesgo.

Factor de riesgo	Categorías	$\beta_j(W_{js} - W_{jREF})$	Puntos$_{js}$= $\beta_j(W_{js} - W_{jREF})/B$
Enfermedad pulmonar			
	No	0	0
	Si	1.224	1
Diabetes			
	No	0	0
	Si	1.296	1
Frecuencia respiratoria			
	<24	0	0
	≥24	2.070	2
Presión arterial sistólica			
	≥95	0	0
	<95	1.659	2
pH			
	≥ 7.35	0	0
	< 7.35	2.197	2

Continúa en la siguiente página.

Factor de riesgo	Categorías	$\beta_j(W_{js} - W_{jREF})$	Puntos$_{js}$= $\beta_j(W_{js} - W_{jREF})/B$
Sodio			
	$\in$[135-145)	0	0
	$\notin$[135-145)	0.966	1
Número de lóbulos afectados			
	1	0	0
	2	0	0
	3 o 4	5.948	6

Nos interesa conocer la probabilidad estimada para que un individuo fallezca ($Y = 1$) dependiendo de la puntuación que ha obtenido en el sistema (de 0 a 15 puntos). El término que necesitamos conocer se aproxima de la siguiente manera:

$$\sum_{j=0}^{q} \beta_j X_j \approx -6.762 + B(\text{total de puntos})$$

Siguiendo esa aproximación, estas son las probabilidades que tiene un enfermo de fallecer dependiendo de la puntuación obtenida en el score ABS.

Puntuación	Probabilidad de mortalidad estimada	Puntuación	Probabilidad de mortalidad estimada
0	0.0016	8	0.7243
1	0.0030	9	0.8735
2	0.0079	10	0.9477
3	0.0206	11	0.9794
4	0.0523	12	0.9921
5	0.1265	13	0.9970
6	0.2757	14	0.9988
7	0.5	15	0.9996

Al igual que en la creación del QDS, debemos tener en cuenta que para que la puntuación se pueda calcular, el individuo no puede tener valores perdidos en ninguna de las 8 variables que forman el ABS, y por tanto, son necesarios tanto un análisis de sangre rutinario como una gasometría, además de otras pruebas.

2.5.3. Validación del ABS

Ahora que conocemos la puntuación que le asigna el score ABS a cada paciente, vamos a comprobar si es buen predictor al aplicarlo a nuevos datos, es decir, vamos a validarlo.
Para validarlo, compararemos las curvas ROC en ambas muestras y miraremos si hay diferencias estadísticamente significativas entre sus AUCs. En

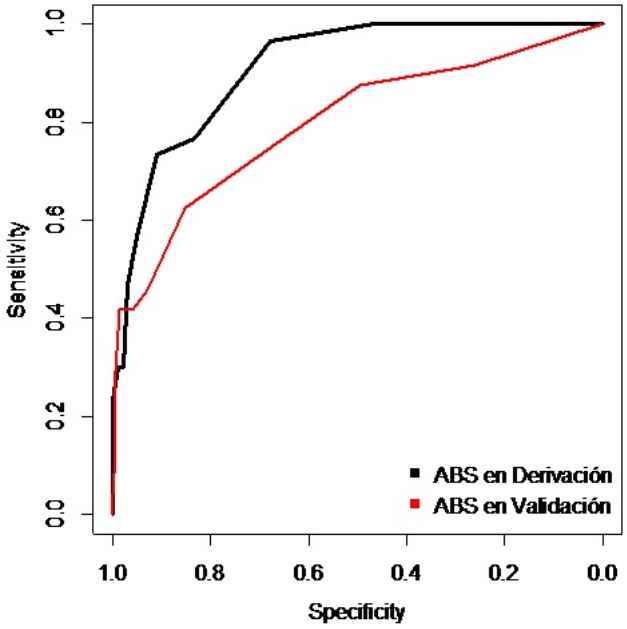

Figura 2.3: Curvas ROC del ABS en derivación y en validación

la Figura 2.3 se aprecia que la curva ROC referente a la muestra de derivación es mejor que la de validación, y por tanto el AUC será mayor. El ABS obtiene un AUC de 0.912 en la muestra de derivación frente al 0.798 de la de validación y al comparar ambas curvas con el test `Dlong` obtenemos un p-valor de 0.054. Como no conseguimos diferencias estadísticamente significativas (p-valor mayor a 0.05) entre ambas muestras y como el AUC consigue buenos resultados en la muestra de validación, asumimos que el score ABS es buen predictor de la mortalidad en otras muestras.

2.5.4. Comparativa con la literatura

Una vez hemos validado este segundo marcador, vamos a comparar los resultados que obtenemos con este score en la muestra total, con los resultados del CURB-65 y del FINE. Partiendo de las variables que indican la puntuación de cada individuo para los tres marcadores, vamos a dibujar las tres curvas ROC y a comparar los AUCs de los scores.

Tal y como se aprecia en la Figura 2.4 la curva ROC del score FINE se acerca más a la curva perfecta, mientras que el CURB-65 obtiene los peores resultados. El AUC del ABS en la muestra completa es de 0.86, el del CURB-65 0.832 y el del FINE 0.894. Al comparar el área bajo la curva ROC de dos

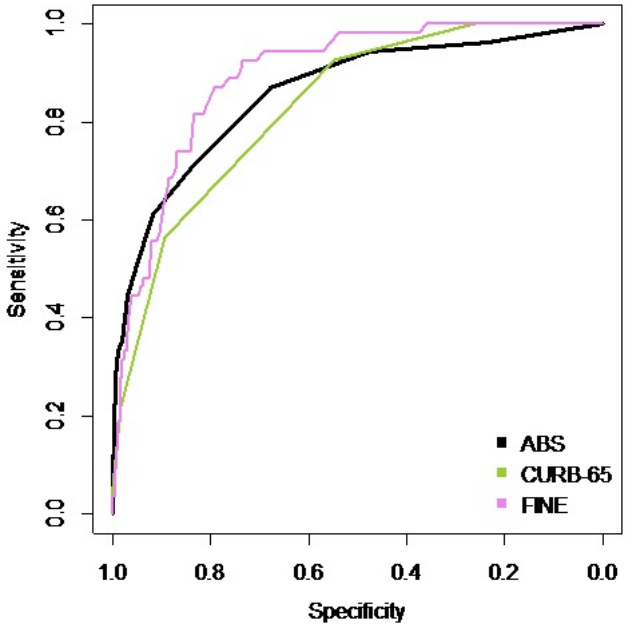

Figura 2.4: Curvas ROC de los tres scores en la muestra completa

en dos, los p-valores que obtenemos son mayores a 0.05 (0.45 al compararlo con el CURB-65 y 0.3161 al compararlo con el FINE), por lo que no hay diferencias estadísticamente significativas entre nuestro segundo score y los otros dos.

Ahora que sabemos que el score ABS es buen predictor de mortalidad al aplicarlo a nuevos datos, y que lo hemos comparado con el CURB-65 y con el FINE, vamos a dividir las puntuaciones en distintos grupos de riesgo. Igual que antes, utilizando la función `catpredi` [10], vamos a dividir el score en 4 categorías distintas dependiendo de la gravedad del paciente: leve, moderado, grave y muy grave.

		No fallece	Fallece
Leve	(0-1 puntos)	502 (99.41 %)	3 (0.59 %)
Moderado	(2 puntos)	212 (98.15 %)	4 (1.85 %)
Grave	(3-4 puntos)	252 (94.84 %)	14 (5.56 %)
Muy grave	(5-15 puntos)	88 (72.73 %)	33 (27.27 %)

Esta tabla muestra cuántos individuos fallecen y cuántos no en cada categoría del score ABS en base a la muestra total. Dependiendo de la puntua-

ción obtenida, clasificamos a los individuos en uno de los siguientes grupos de riesgo:

- **Grupo I**: Los pacientes de esta categoría tienen buen pronóstico y pueden tratarse en domicilio.
- **Grupo II**: Si el individuo obtiene 2 puntos en el sistema de puntuación debe valorarse una estancia corta en el hospital o tratamiento en domicilio con supervisión médica.
- **Grupo III**: Los pacientes de esta categoría necesitan ingreso hospitalario.
- **Grupo IV**: En estos pacientes se debe valorar el ingreso en UCI.

Hemos observado que el score ABS categorizado también tiene buena capacidad discriminante siendo el AUC estimado de 0.8482.

Capítulo 3

Conclusiones

En este trabajo hemos creado dos scores que relacionan la variable respuesta mortalidad por neumonía con el resto de factores de riesgo. Parecía interesante comparar nuestros scores con otros dos que utilizan los neumólogos, el CURB-65 y el FINE. Tal y como hemos dicho, el CURB-65 es un score más simple que el FINE y por eso, creímos conveniente crear dos marcadores: el *Quick-Decision Score* (QDS) y *Analitic-Based Score* (ABS). Para poder dar valor a los factores de riesgo que forman el score QDS bastará con una rápida valoración clínico-radiológica, mientras que para el ABS necesitaremos además realizar pruebas analíticas.

Hemos utilizado el método *split-sample validation* para validar nuestros scores. En el score QDS se aprecian diferencias estadísticamente significativas entre la muestra de derivación y de validación (p-valor 0.044), mientras que con el score ABS no (p-valor 0.054). A pesar de obtener estos resultados, el AUC en la muestra de validación ronda el 0.8 para los dos scores, y por tanto, podemos considerar que la capacidad discriminativa de ambos scores es satisfactoria.

En base a los resultados obtenidos y el tamaño muestral disponible, consideramos que sería interesante desarrollar los modelos en la muestra total y utilizar otra técnica para la validación de los scores como la técnica del *bootstrap*, ya que esta no pierde información de los pacientes en ninguna de las submuestras. Lo ideal hubiera sido conseguir una segunda muestra formada por enfermos de neumonía y hacer lo que se conoce como una validación externa, pero no fue posible.

Al comparar el CURB-65 con el QDS (el simple), observamos que no hay diferencias estadísticamente significativas entre el AUC de la curva ROC de ambos scores en la muestra total (p-valor de 0.379), y por tanto, hemos creado un nuevo score bastante simple que obtiene muy buenos resultados.

Por otro lado, al comparar el score ABS con el FINE, concluimos diciendo que no hay diferencias estadísticamente significativas entre el AUC de ambos

scores en la muestra total (p-valor 0.3161) y por tanto, hemos creado y validado un score bastante más simple que el FINE que tiene buena capacidad discriminativa.

En conclusión, hemos logrado dos scores aplicables a pacientes con diagnóstico de neumonía, que ayudan al médico a tomar desiciones de manera rápida y eficaz. Dado que hemos validado el score ABS y que es bastante más simple que el FINE, pensamos que nuestro score ABS puede ser de gran aplicabilidad clínica.

Bibliografía

[1] Hosmer D.W. and Lemeshow S., Applied Logistic Regression, (2000).

[2] Mc Cullagh and Nelder, Generalized Linear Models, (1989).

[3] Sullivan L.M. et al, "Presentation of multivariate data for clinical use: The Framingham Study risk score functions", Stadistics in Medicine, Volumen 23 (2004), páginas 1631-1660.

[4] Steyerberg E.W., Clinical Prediction Models: A Practical Approach to Development, Validation and Updating, (2008), páginas 299-312.

[5] Lim W.S. et al, "Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study", Thorax número 58 (2003), páginas 377-382.

[6] Fine M.J. et al, "A prediction rule to identify low-risk patients with community-acquired pneumonia", The New England Journal of Medicine, Volumen 336 número 4 (1997), páginas 243-250.

[7] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing 2013. URL http://www.R-project.org/.

[8] Rovin X., Turck N. et al, "pROC: an open-source package for R and S+ to analyze and compare ROC curves", BCM Bioinformatics, Volumen 12 (2011), página 77.

[9] DeLong E.R., DeLong D.M. and Clarke-Pearson D.L., "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach", Biometrics, Volumen 44 (1988), páginas 837-845.

[10] Barrio I., Arostegui I., Rodriguez-Alvarez M.X. and Quintana J.M. (2015). A new approach to categorising continuous variables in prediction models: proposal and validation. Technical report.