MODELOS DE ELECCIÓN DISCRETA LOGIT Y PROBIT. REGRESIÓN DE COX

MODELOS CON VARIABLES CUALITATIVAS: MODELOS DE ELECCIÓN DISCRETA

Aunque el rango de valores posibles de una variable estadística es usualmente un intervalo de la recta real, sin embargo la teoría econométrica considera también modelos de regresión lineal en los que alguna de las variables explicativas toma valores en un conjunto discreto y finito. Un ejemplo de este tipo de variables lo constituye las llamadas *variables ficticias* muy utilizadas en la teoría de modelos y que suelen introducirse tradicionalmente para incorporar al modelo posibles variaciones estructurales ocurridas durante el período muestral, o efectos socioeconómicos que pudieran distinguir el comportamiento de unos individuos de otros.

En estos casos, se introduce en el modelo una variable discreta para cada una de las características que se pretenden tomar en consideración. A cada una de las posibles modalidades que puede presentar la característica se le asocia un valor numérico, y la *variable ficticia* así definida, se utiliza en la estimación del modelo como una variable explicativa más. Si se pretende recoger en el modelo la idea de un posible cambio estructural en el instante t_0 , entonces la variable ficticia correspondiente tomará el valor 0 para $1 < t < t_0$, y el valor 1 para $t_0 < t < T$. Si se pretendiese discriminar entre los gastos en educación de las familias de una muestra dependiendo de su pertenencia a un medio rural o urbano, podría definirse una variable ficticia que toma el valor 0 si la familia vive en un medio rural, y el valor 1 si vive en un medio urbano.

Hay otras variables *discretas* que podrían ser influyentes en el comportamiento de la variable endógena sin ser variables ficticias. Por ejemplo, el número de hijos es importante al analizar el gasto en educación de una determinada familia. Conviene recordar que variables continuas como la edad o los ingresos de un individuo, se transforman en ocasiones en variables discretas debido al diseño de la muestra usada en la recogida de datos.

Por otra parte, son muchos los problemas y cuestiones de interés en economía, demografía, sociología, epidemiología, biología, medicina y otras ciencias en los que es la variable endógena la que puede no ser continua. A veces, ni siquiera esta variable es cuantificable, como ocurre cuando se pretende cuantificar el tipo de transporte que cada persona de un determinado colectivo toma diariamente cuando acude a su trabajo. En tal caso pueden considerarse como alternativas la utilización de coche particular o de transporte público, pudiendo introducir la variable endógena que vale 0 o 1 según que el medio de transporte utilizado sea privado o público.

Otros ejemplos de variable endógena (dependiente) cualitativa podría ser un modelo que trate de explicar el nivel de estudios de los individuos en función de sus características personales, o un modelo que tratase de explicar el votar o no a un determinado partido político en elecciones según determinadas variables que influyen en ello, el tipo de escuela (privada o pública) a la que envían las familias a los hijos, en función de variables como los ingresos, la edad, etc., la devolución o no en fecha de vencimiento de un crédito concedido por una entidad bancaria en función de determinadas características de los deudores.

Los modelos de regresión más comunes exigen que la distribución de la variable dependiente sea normal. Cuando dicha variable no es normal pero sí continua, se pueden intentar, para poder aplicar dichos modelos, transformaciones de la misma que normalicen su función de probabilidad. Pero esta estrategia, sin embargo, no sirve para variables discretas. En este Capítulo se va a introducir una ampliación en la teoría de modelos de regresión consistente en estudiar un modelo de regresión aplicable a variables dependientes discretas binomiales.

Se dice que un proceso es binomial cuando sólo tiene dos posibles resultados: "éxito" y "fracaso", siendo la probabilidad de cada uno de ellos constante a lo largo de una serie de repeticiones. A la variable número de éxitos en n repeticiones se le denomina variable binomial. A la variable resultado de un sólo ensayo y , por tanto, con sólo dos valores: 0 para fracaso y 1 para éxito, se le denomina *binomial puntual* o de *Bernouilli*. La terminología de "éxito" y "fracaso" proviene de los juegos de azar (una apuesta en un juego de azar sólo tiene dos resultados, se gana o se pierde) y se mantiene, aunque no es muy afortunada en otros contextos. La presencia o ausencia de un cierto carácter biológico, por ejemplo, que una pareja tenga un descendiente con ojos negros, es un fenómeno binomial (sólo dos resultados y la misma probabilidad en sucesivas repeticiones) sin que haya ninguna connotación de éxito en dicha presencia, incluso en Ciencias de la Salud y para el proceso, también binomial, de padecer, o no, una cierta enfermedad parece claramente inapropiado denominar éxito a enfermar.

Recuérdese que la función densidad de probabilidad de una variable binomial Y de parámetros n y p es la siguiente:

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y}$$

La función de densidad para una variable binomial puntual (o variable de Bernouilli) viene dada por:

$$f(y) = p^y (1 - p)^{1-y}$$

Un proceso binomial puntual está, pues, caracterizado por la probabilidad de éxito, representada por p (es el único parámetro de su función de probabilidad), y por la probabilidad de fracaso que se representa por q . Evidentemente, ambas probabilidades están relacionadas por $p+q=1$. En ocasiones, se usa el cociente p/q , denominado "*odds*" (o "*ventaja*") que indica cuánto más o menos probable es el éxito que el fracaso. El *odds* se define como:

$$\text{Odds} = \text{Probabilidad de éxito} / \text{Probabilidad de fracaso} = p/q$$

Por ejemplo; si la probabilidad de éxito es 0,75, la de fracaso será 0,25, con lo que el *odds* valdrá $0,75/0,25=3$. Esto puede interpretarse diciendo que el éxito es tres veces más probable que el fracaso. Siempre que el *odds* sea mayor que 1 la probabilidad de éxito será mayor que la fracaso, si el *odds* es menor que el fracaso es más probable que el éxito y si el *odds* es la unidad éxito y fracaso son igualmente probables.

También es posible obtener la probabilidad de éxito conociendo el *odds*, ya que se tiene lo siguiente:

$$\text{Odds} = p/q = p/(1-p) \Rightarrow p = \text{odds}/(\text{odds}+1)$$

Por ejemplo; si el *odds* vale 3, la probabilidad de éxito es $p=3/4=0.75$.

Otro concepto, que se utiliza para comparar el *odds* de dos sucesos es el "*odds ratio*" (o "*ventaja comparativa*"), que se define como el cociente de los *odds* para los dos sucesos, y por lo tanto mide cuánto más ventajoso es el éxito sobre el fracaso en el primer suceso que en el segundo.

Suele denotarse el *odds* para un suceso por la letra griega π , mientras que el *odds ratio* de dos sucesos se denota por la letra griega θ . Si $\theta = \pi_1/\pi_2$ es la unidad, los *odds* para los dos sucesos son iguales, con lo que las probabilidades de éxito en los dos sucesos serán iguales, ya que $p = \text{odds}/(\text{odds}+1)$ vale lo mismo para los dos sucesos. Si $\theta = \pi_1/\pi_2 > 1$ el éxito es más ventajoso sobre el fracaso en el primer susceso. Si $\theta = \pi_1/\pi_2 < 1$ el éxito es más ventajoso sobre el fracaso en el segundo susceso. El *odds ratio* se denomina también "*riesgo relativo*".

EL MODELO DE REGRESIÓN LOGÍSTICA

Un modelo de regresión con variable dependiente binomial (modelo logístico o modelo de regresión logística) será un modelo que permita estudiar si dicha variable discreta depende o no, de otra u otras variables. Si una variable binomial de parámetro p es independiente de otra variable X , se cumple $(p|X=x) = p$, para cualquier valor x de la variable X .

Por consiguiente, un modelo de regresión con variable dependiente binomial y una única variable independiente X se materializa en una función en la que p aparece dependiendo de X y de unos coeficientes cuya investigación permite abordar la relación de dependencia. Para una única variable independiente X , el modelo de regresión logística toma la forma:

$$\ln(p/q | X = x) = \alpha_0 + \alpha_1 X, \text{ o de forma simplificada: } \ln(p/q) = \alpha_0 + \alpha_1 X$$

donde $\ln$ significa logaritmo neperiano, α_0 y α_1 son constantes y X una variable que puede ser aleatoria o no, continua o discreta. Este modelo se puede fácilmente generalizar para k variables independientes, dando lugar al *modelo logístico múltiple*, que se expresa como sigue:

$$\ln(p/q) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_k X_k$$

Hemos definido el *modelo logístico* como el logaritmo del *odds* para el suceso que representa la variable aleatoria binomial puntual dependiente del modelo.

Hay varias razones para plantear el modelo con el logaritmo de *odds*, en lugar de plantearlo simplemente con la probabilidad de éxito o con el *odds*. En primer lugar, el campo de variación de $\ln(p/q)$ es todo el campo real (de $-\infty$ a ∞), mientras que, para p el campo es sólo de 0 a 1 y para p/q es de 0 a ∞ . Por lo tanto, con el modelo definido en función del logaritmo de *odds* no hay que poner restricciones a los coeficientes que complicarían su estimación. Por otro lado, y más importante, en los modelos en función del logaritmo de *odds* los coeficientes son fácilmente interpretables en términos de independencia o asociación entre las variables, como se verá más adelante.

El modelo logístico se puede escribir de otras formas equivalentes que para ciertas aplicaciones son más cómodas de manejar. Tenemos:

$$\begin{aligned} \ln(p/q) = \alpha_0 + \alpha_1 X &\Leftrightarrow \ln\left(\frac{p}{1-p}\right) = \alpha_0 + \alpha_1 X \Leftrightarrow \frac{p}{1-p} = e^{\alpha_0 + \alpha_1 X} \Leftrightarrow \\ p &= \frac{e^{\alpha_0 + \alpha_1 X}}{1 + e^{\alpha_0 + \alpha_1 X}} \Leftrightarrow p = \frac{1}{1 + e^{-(\alpha_0 + \alpha_1 X)}} \end{aligned}$$

Estas dos últimas expresiones, si son conocidos los coeficientes α_0 y α_1 , permiten calcular directamente la probabilidad del proceso binomial para los distintos valores de la variable X .

A la función: $f(z) = \frac{1}{1 + e^{-z}}$ se le denomina función logística. El modelo de regresión logística modeliza la probabilidad de un proceso binomial como la función logística de una combinación lineal de la variable dependiente.

El modelo de regresión logística múltiple tendrá la expresión:

$$p = \frac{1}{1 + e^{-(\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \dots + \alpha_k X_k)}}$$

INTERPRETACIÓN DE LOS COEFICIENTES DEL MODELO

Cuando $X=0$ el modelo logístico $\ln(p/q) = \alpha_0 + \alpha_1 X$ queda de la forma: $\ln(p/q | X = 0) = \alpha_0$, con lo que ya podemos decir que α_0 es el **logaritmo de odds cuando la variable independiente es cero**.

Cuando $X=1$ y la variable X sólo puede tomar los valores 0 y 1 se tiene que:

$$\ln(p/q | X = 1) = \alpha_0 + \alpha_1 = \ln(p/q | X = 0) + \alpha_1 \Leftrightarrow$$

$$\alpha_1 = \ln(p/q | X = 1) - \ln(p/q | X = 0) = \ln \frac{p/q | X = 1}{p/q | X = 0}$$

Por lo tanto, se puede decir que α_1 es el **logaritmo del cociente de los odds para los valores de la variable X , supuesto que X sólo pueda tomar los valores 0 y 1**. Por lo tanto, α_1 es el **odds ratio para los dos únicos valores 0 y 1 de la variable X** .

Además, si la variable binomial es independiente la variable X , ambos *odds* son iguales, por lo tanto el *odds ratio* es 1 y su logaritmo es cero. Por lo tanto, **para estudiar con un modelo logístico la independencia de las variables basta comprobar si el coeficiente α_1 es cero**.

Decir que α_0 es el logaritmo de *odds* cuando la variable independiente es cero es lo mismo que decir que e^{α_0} es el *odds* cuando $X=0$.

Decir que α_1 es el logaritmo del cociente de los *odds* para los dos valores de la variable X es lo mismo que decir e^{α_1} es el *odds ratio* entre $X=1$ y $X=0$.

Si la variable X puede tomar más de dos valores, evidentemente se sigue manteniendo que e^{α_1} es el odds cuando $X=0$. Además, en este caso, e^{α_1} es el odds ratio para el aumento en una unidad en la variable X (lo que implica que el odds ratio para el aumento en una unidad en la variable X es constante en el modelo de regresión logística).

Cualesquiera que sean los valores que tome la variable X se cumple que: $\ln(p/q | X = x_0) = \alpha_0 + \alpha_1 x_0$ y $\ln(p/q | X = x_1) = \alpha_0 + \alpha_1 x_1$, con lo que:

$$\underbrace{\ln(p/q | X = x_1) - \ln(p/q | X = x_0)} = (\alpha_0 + \alpha_1 x_1) - (\alpha_0 + \alpha_1 x_0) = \alpha_1(x_1 - x_0) = \alpha_1 \delta$$

$$\ln \frac{p/q | X = x_1}{p/q | X = x_0} = \alpha_1 \delta \Rightarrow \text{odds ratio} = e^{\alpha_1 \delta} = (e^{\alpha_1})^\delta$$

Por lo tanto, para un aumento de la variable X desde x_0 a x_1 (es decir, para un aumento en X de valor $\delta=x_1-x_0$) se tiene que $e^{\alpha_1 \delta}$ es el odds ratio para el aumento en δ unidades en la variable X. Cuando $\delta=1$, se tiene que el odds ratio para el aumento en una unidad en la variable X es constante e igual a e^{α_1} en el modelo de regresión logística, es decir, que *el logaritmo del odds ratio para el aumento en una unidad en la variable X es constante y precisamente igual al coeficiente α_1 del modelo.*

ESTIMACIÓN DE LOS COEFICIENTES

El método de los mínimos cuadrados, clásico en la estimación de los coeficientes de los modelos de regresión no es aplicable al modelo logístico, ya que dicho método se basa en la normalidad de la variable dependiente, que en este caso no se cumple. Por otra parte, cuando $q=0$, es imposible calcular $\ln(p/q)$. Se tratará entonces de utilizar el método de máxima verosimilitud.

Comenzaremos considerando el caso más simple con una sola variable independiente X. Tomamos una muestra de n observaciones (y_i, x_i) para la variable puntual binomial dependiente Y y para la variable independiente X. La variable Y toma valores y_i que sólo pueden ser 1 con probabilidad p_i o 0 con probabilidad $1-p_i$. Como x_i depende de p_i a través del modelo logístico tenemos:

$$p_i = \frac{e^{\alpha_0 + \alpha_1 x_i}}{1 + e^{\alpha_0 + \alpha_1 x_i}}$$

La función de verosimilitud para una variable binomial puntual es:

$$L(p_i | y_i) = (p_i)^{y_i} (1 - p_i)^{1 - y_i}$$

y para n observaciones independientes la función de verosimilitud de la muestra será:

$$L(p_1, \dots, p_n | y_1, \dots, y_n) = \prod_{i=1}^n (p_i)^{y_i} (1 - p_i)^{1-y_i}$$

y al representar p_i por el modelo logístico tendremos ya la expresión de la función de verosimilitud para la muestra como función de los parámetros a estimar:

$$\begin{aligned} L(p_1, \dots, p_n | y_1, \dots, y_n) &= \prod_{i=1}^n \left(\frac{e^{\alpha_0 + \alpha_1 X_i}}{1 + e^{\alpha_0 + \alpha_1 X_i}} \right)^{y_i} \left(1 - \frac{e^{\alpha_0 + \alpha_1 X_i}}{1 + e^{\alpha_0 + \alpha_1 X_i}} \right)^{1-y_i} = \\ &= \prod_{i=1}^n \left(\frac{e^{\alpha_0 + \alpha_1 X_i}}{1 + e^{\alpha_0 + \alpha_1 X_i}} \right)^{y_i} \left(\frac{1}{1 + e^{\alpha_0 + \alpha_1 X_i}} \right)^{1-y_i} = \frac{e^{\alpha_0 \sum_{i=1}^n y_i + \alpha_1 \sum_{i=1}^n X_i y_i}}{\prod_{i=1}^n (1 + e^{\alpha_0 + \alpha_1 X_i})} = L(\alpha_0, \alpha_1) \end{aligned}$$

y como suele ser usual en máxima verosimilitud, maximizaremos el logaritmo de la función $L(\alpha_0, \alpha_1)$ en vez de la función misma. Los parámetros estimados del modelo serán los valores de α_0 y α_1 , que maximicen la función $\ln L(\alpha_0, \alpha_1)$.

ESTIMACIÓN POR INTERVALOS Y CONTRASTES DE HIPÓTESIS SOBRE LOS COEFICIENTES

Según el teorema central del límite, los estimadores por máxima verosimilitud de los parámetros del modelo logístico son asintóticamente normales y su matriz de varianzas covarianzas es perfectamente calculable a partir del algoritmo de maximización de la función de verosimilitud (método de Newton Rhampson).

De esta forma, un intervalo de confianza al $(1 - \alpha)\%$ para el estimador del coeficiente α_i del modelo será:

$$\hat{\alpha}_i \pm Z_{\alpha/2} \hat{\sigma}(\hat{\alpha}_i)$$

Hay que tener presente que los estimadores habituales que miden asociación entre variables son los *odds ratio*, por lo tanto interesa dar los intervalos de confianza para los *odds ratio*, que evidentemente serán:

$$e^{\hat{\alpha}_i \pm Z_{\alpha/2} \hat{\sigma}(\hat{\alpha}_i)}$$

El estadístico para el contraste:

$$H_0: \alpha_i = a$$

$$H_1: \alpha_i \neq a$$

Será: $Z = \frac{\hat{\alpha}_i - a}{\hat{\sigma}(\hat{\alpha}_i)} \rightarrow N(a, \hat{\sigma}(\hat{\alpha}_i))$ y región crítica $|Z| > Z_{\alpha/2}$

También suele utilizarse para el contraste el estadístico de Wald definido como $W=Z^2$ y cuya distribución es una chi-cuadrado con 1 grado de libertad. La región crítica de este contraste es $W > \chi^2_{\alpha}$.

En el modelo logístico es muy interesante contrastar la hipótesis $\alpha_i = 0$ para $i=1, \dots, k$, porque no rechazar esta hipótesis para un valor de i implica que la variable Y no depende X_i , y por lo tanto esta última no debería figurar en el modelo.

También suele utilizarse el contraste de la razón de verosimilitudes, basado en el estadístico $-2\text{Log}(L_0/L_1)$ donde L_0 es el máximo de la función de verosimilitud bajo la hipótesis nula y L_1 es el máximo de la función de verosimilitud bajo la hipótesis alternativa. Este estadístico tiene una distribución gi-cuadrado con grados de libertad igual al número de parámetros bajo la hipótesis nula.

MODELOS PROBIT Y LOGIT

Como los valores de una probabilidad están entre cero y uno, las predicciones realizadas con los modelos de elección discreta deben estar acotadas para que caigan en el rango entre cero y uno. El modelo general que cumple esta condición se denomina **modelo lineal de probabilidad**, y tiene la forma funcional:

$$P_i = F(x_i, \beta) + u_i$$

Se observa que si F es la función de distribución de una variable aleatoria entonces P varía entre cero y uno.

En el caso particular en que la función F es la función logística estaremos ante el **modelo Logit**, cuya forma funcional será la siguiente:

$$P_i = F(x_i, \beta) + u_i = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}} + u_i$$

En el caso particular en que la función F es la función de distribución de una normal unitaria estaremos ante el **modelo Probit**, cuya forma funcional será la siguiente:

$$P_i = F(x_i, \beta) + u_i = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{x_i\beta} e^{-\frac{t^2}{2}} dt + u_i$$

ANÁLISIS DE LA SUPERVIVENCIA

En los últimos años se han popularizado los modelos de probabilidad que intentan analizar, bien las series de tiempos de fallo en un proceso industrial, bien las series de tiempos de supervivencia de un grupo de individuos, bien las series de duraciones de ciertos procesos, bien las series de tiempos que se permanece empleado en la misma empresa, de tiempos de permanencia en la universidad, de tiempos desde el matrimonio a la llegada del primer hijo, etc. Tiene cierto interés conocer los años de supervivencia tras una intervención quirúrgica (un trasplante de corazón, por ejemplo) o tras el diagnóstico de una enfermedad (el SIDA u otras). Quizás se podría pretender determinar el tiempo que transcurre desde la administración de determinado fármaco y la desaparición de sus efectos. Obsérvese que todos los ejemplos se refieren a un suceso único, irrepetible sobre los individuos y no recurrente. El análisis de este tipo de series, que normalmente siempre son series de tiempo, suele conocerse con el nombre genérico de *análisis de la supervivencia* (bien sea referida a individuos, máquinas, etc.).

La variable de interés en el análisis de supervivencia es la longitud del período de tiempo que transcurre desde el principio de algún acontecimiento hasta el final del mismo, o hasta el momento en que ese acontecimiento es observado, lo que puede ocurrir antes de que el acontecimiento acabe. Los datos habitualmente se presentan como un conjunto de duraciones o supervivencias, $t_1, t_2, \dots, t_n$ que no necesariamente tienen porqué empezar en el mismo punto del tiempo.

Una característica inherente al análisis de supervivencia es la *censura*. Se dice que *los datos están censurados* si no se pueden observar por completo. Considerese por ejemplo el análisis del tiempo que transcurre entre el diagnóstico de un determinado tipo de cáncer en un grupo de pacientes y la muerte de los mismos. Los pacientes son observados cada seis meses, empezando justo en el momento en que se les diagnosticó el cáncer. Por el momento supóngase que a todos los pacientes se les diagnosticó el cáncer el mismo día. Tras seis meses algunos pacientes han muerto y otros no. Para los pacientes sobrevivientes la duración, o supervivencia, es por lo menos igual al período observado, $t_i = 6$ meses, pero no es igual a él. Este tipo de censura, la más habitual, se conoce como *censura por la derecha* (tiempo de supervivencia real mayor que el observado).

Es posible, así mismo, que exista *censura por la izquierda*, en cuyo caso el tiempo de supervivencia real es menor que el observado. Supongamos por ejemplo que estamos interesados en la supervivencia de un grupo de pacientes con síntomas de un determinado tipo de cáncer, hayan sido o no diagnosticados. En este caso algunos pacientes pueden haber muerto antes de que se les diagnosticase. Tales pacientes presentan censura por la izquierda.

La censura también puede ser *censura de intervalo*, por cuanto se conoce que el evento irrepitable ha ocurrido en un intervalo de tiempo determinado. Supongamos ahora que algunos de los pacientes sobrevivientes seis meses después de serles diagnosticado el cáncer, han muerto en la observación, un año después. Existe entonces una censura de intervalo, entre seis meses y año.

TABLAS DE MORTALIDAD

En el análisis de la supervivencia lo primero que se analiza es la tabla de mortalidad (o tabla de vida). Esta tabla contiene distintas informaciones descriptivas sobre la evolución de las observaciones, entre las que tenemos las siguientes:

- **Intervalos (Intervals):** Aparecen los límites inferior y superior de los intervalos de tiempo en que se ha dividido la serie, y en cada uno de los cuales se han registrado el correspondiente número de fallos (fallecimientos, averías, etc.).
- **Número de fallos en cada intervalo (Number of Failures).**
- **Número de abandonos (Number Withdrawn):** Individuos que no han llegado al final de periodo debido a abandono (mortalidad experimental) y no a causa de producirse el evento terminal (fallecimiento). Se puede observar cómo puede haber observaciones que no han finalizado el primer intervalo y sin embargo no han fallecido. También hay individuos sobre los que no se tiene información completa para ciertos intervalos de tiempo. Por tanto, no se pueden contabilizar como fallecimientos, ni como vivos, sino como abandonos.
- **Número de observaciones expuestas a riesgo (Number at Risk).**
- **Funciones de supervivencia:** En las tablas de vida se incluyen funciones de supervivencia (*survival functions*) con objeto de facilitar la interpretación de los resultados. Las principales funciones de supervivencia son: la **función de supervivencia acumulada** estimada, la **función de riesgo** (o azar) estimada y la **función de densidad** estimada para la distribución de los tiempos de la serie de vida en estudio.

Más formalmente, sea T una variable aleatoria continua no negativa con función de densidad $f(t)$, que representa el tiempo de supervivencia (por ejemplo, de un paciente, de una máquina, etc.). Su función de distribución, o función de probabilidad acumulada es $F(t) = \text{Prob}(T \leq t)$. La función de supervivencia $S(t)$ se define como la probabilidad de supervivencia hasta t , o sea $S(t) = \text{Prob}(T \geq t) = 1 - F(t)$. La función de riesgo o tasa de azar $h(t)$ se define como la probabilidad de que un individuo, que se supone vivo en el instante t , sobreviva al siguiente intervalo de tiempo lo suficientemente pequeño, o sea, $h(t)$ es la función de densidad condicional en t dada la supervivencia hasta t y se tiene $h(t) = f(t)/S(t)$.

La función de densidad, la función de riesgo y la función de supervivencia están relacionadas mediante $f(t) = S(t)h(t)$ y $h(t) = -d\ln S(t)/dt$. Otra función de interés es la función integrada o acumulada de riesgo $H(t) = -\ln S(t)$. En ocasiones se dice también que la variable T representa el tiempo de fallo, sobre todo en teoría de la fiabilidad.

El objetivo del análisis de supervivencia es estimar las funciones de supervivencia y de riesgo a partir de los tiempos de supervivencia observados. Existen dos métodos principales para el cálculo de estas funciones: a) método actuarial de Berkson y Gage (1950) y b) método del producto de Kaplan y Meyer (1958). Una síntesis de estos métodos puede verse en Pardell, Cobo y Canela (1986).

Estimaciones no paramétricas de la función de supervivencia

Este procedimiento más común que permite realizar estimaciones no paramétricas de la función de supervivencia, basadas en la función de supervivencia empírica de la muestra, es el ***método del límite producto de Kaplan Meier*** (la función de supervivencia empírica es $S_m(t) = \text{Número de individuos con tiempo de supervivencia mayor o igual que } t \text{ dividido entre el número total de individuos}$, y se tiene que $S_m(t) = 1 - F_m(t)$ donde $F_m(t)$ es la función de distribución empírica). Se usa para obtener probabilidades de supervivencia para datos multicensados y también se usa en ensayos clínicos para estudiar y comparar tasas de supervivencia de pacientes bajo diferentes tratamientos.

Estimaciones paramétricas de la función de supervivencia

Las aproximaciones no paramétricas no necesitaban especificar tipo de distribución de probabilidad para los tiempos de supervivencia. De este modo, la función de riesgo tampoco necesita ser especificada permitiendo, por tanto, una gran flexibilidad en el análisis. Ahora bien, cuando los datos respondan efectivamente a una determinada distribución de probabilidad, las inferencias basadas en la parametrización de dicha distribución serán más precisas o eficientes. Si la distribución de probabilidad asumida es correcta, los errores estándar de los estimadores en las aproximaciones paramétricas son menores. Además estas aproximaciones permiten realizar inferencias poblacionales no limitándose a la muestra analizada como en el caso de las alternativas puramente no paramétricas.

Supongamos ahora que los datos siguen un modelo de probabilidad determinado. El modelo más sencillo es el que supone que la tasa de riesgo no varía en el tiempo, es decir $h(t)$ es constante. En este caso, la probabilidad condicionada a estar vivo en t de que un individuo muera en un intervalo de tiempo determinado (lo suficientemente pequeño) será la misma con independencia del momento en el que se observe el individuo.

Esta característica se conoce como ***pérdida de memoria***. La función de riesgo puede representarse por $h(t) = \lambda$ para $0 \leq t \leq \infty$ siendo λ una constante positiva. Como $-d\ln S(t)/dt = h(t) = \lambda \Rightarrow S(t) = Ke^{-\lambda t}$, y como $S(0) = 1$ entonces $K = 1$, con lo que $S(t) = e^{-\lambda t}$ y estamos ante la ***distribución exponencial*** para los datos, porque la función de densidad es $f(t) = h(t)S(t) = \lambda e^{-\lambda t}$.

El problema de la distribución exponencial es que, salvo en procesos industriales, es difícilmente sostenible que la supervivencia se defina por una tasa de riesgo constante. Por este motivo se han propuesto otras distribuciones alternativas, entre las cuales, la más utilizada es la *distribución de Weibull*, que supone que la tasa de riesgo toma la forma $h(t) = \lambda\gamma t^{\gamma-1}$ para $0 \leq t \leq \infty$ y donde los parámetros λ (*parámetro de escala*) y γ (*parámetro de forma*) son constantes positivas. Si $\gamma = 1$, la función de riesgo es constante, con lo que los tiempos de supervivencia siguen una distribución exponencial. Para otros valores de γ , la función de riesgo crece o decrece de forma monótona (no cambia de dirección). Para el valor ya conocido de $h(t)$ tenemos:

$$S(t) = e^{-\lambda t^\gamma} \Rightarrow f(t) = \lambda\gamma t^{\gamma-1} e^{-\lambda t^\gamma} \quad (\text{función de densidad de la variable Weibull}).$$

Existen otros modelos típicos en el análisis de la supervivencia, como por ejemplo, el *modelo log-logístico*, cuya función de riesgo es $h(t) = (e^\theta \kappa t^{\kappa-1}) / (1 + e^\theta t^\kappa)$, siendo la función de supervivencia $S(t) = (1 + e^\theta t^\kappa)^{-1}$ y siendo la función de densidad $f(t) = (e^\theta \kappa t^{\kappa-1}) / (1 + e^\theta t^\kappa)^2$, que es la función de densidad de una variable log-logística.

REGRESIÓN DE COX

Del mismo modo que las tablas de mortalidad y el análisis de supervivencia de Kaplan-Meier, la regresión de Cox es un método para crear modelos para datos de tiempos de espera hasta un evento con casos censurados presentes. Sin embargo, la regresión de Cox permite incluir en los modelos variables predictoras (covariables). La regresión de Cox gestionará los casos censurados correctamente y proporcionará las estimaciones de los coeficientes para cada una de las covariables, permitiendo evaluar el impacto de múltiples covariables en el mismo modelo. Además, es posible utilizar la regresión de Cox para examinar el efecto de covariables continuas.

La función de azar para supervivencia en el tiempo de los miembros de una población viene dada según Cox por:

$$H(t|x) = h_0(t) \exp(\beta x(t))$$

donde $x(t)$ es un vector de covariables posiblemente dependientes del tiempo y β es un vector de parámetros de regresión a estimar. La función de supervivencia será:

$$S(t|x) = S_0(t) \exp(\beta x(t)) \quad \text{dónde} \quad S_0(t) = \int_0^t h_0(u) du$$

La regresión de Cox permite contrastes de hipótesis lineales a cerca de los parámetros de la regresión, cálculo de intervalos de confianza y análisis de los residuos.