# Tema 5

# Modelos de regresión

Este tema está dedicado al estudio de métodos de modelización de tiempos de supervivencia incorporando la información proporcionada por lo que llamaremos un vector de variables explicativas. Un modelo de covariables a veces explica y/o predice por qué determinadas unidades fallan rápidamente y otras sobreviven un largo periodo de tiempo.

## *5.1. Modelos de regresión en fiabilidad y supervivencia*

Entre estos modelos los de uso más extendido en estudios de supervivencia son *el modelo de tiempo de vida acelerada* y *el modelo de riesgos proporcionales*. En especial este último.

Sea $\mathbf{z} = (z_1, z_2, \ldots, z_k)'$ un vector de dimensión k que contiene k variables asociadas a un sujeto particular. En estudios de fiabilidad y supervivencia posibles variables explicativas incluyen (dependiendo claro está del contexto):

- Variables continuas: tensión, temperatura, voltaje, presión, nivel de colesterol;
- Variables discretas: número de usuarios simultáneos de un sistema, edad;
- Variables categóricas: fabricante, diseño, localización, tratamiento, sexo.

Atendiendo a las distintas áreas de aplicación, podríamos encontrarnos con distintas situaciones, por ejemplo:

- Fiabilidad:
  T = tiempo de fallo de un mecanismo eléctrico;
  $z_1$ = número de veces que se usa/unidad de tiempo;
  $z_2$ = fuente de alimentación eléctrica;
  $z_3$ = condiciones del entorno de uso (temperatura, presión…).
- Bioestadística:
  T = tiempo de supervivencia de un paciente;
  $z_1$ = edad;
  $z_2$ = sexo;
  $z_3$ = nivel de colesterol;
  $z_4$ = tratamiento.
- Sociología:
  T = tiempo de reingreso en prisión de un individuo;
  $z_1$ = edad;
  $z_2$ = tiempo cumplido de condena;
  $z_3$ = número de condenas previas.

La idea general de un modelo de regresión de este tipo, consiste en expresar la distribución del tiempo de fallo como una función de k variables explicativas agrupadas en el vector **z**.

Pueden sugerirse distintos modelos de regresión, a partir de teorías físico–químicas, ajuste de curvas a observaciones empíricas y diversas combinaciones de teoría y práctica. Una importante clase de modelos de regresión supone que uno o más parámetros $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)$, del modelo es una función de las variables explicativas.

Generalmente se emplea una función con una forma específica con uno o más parámetros a estimar de los datos. Esto es una generalización del análisis de regresión estadístico en el que los modelos más usados consisten en una distribución normal cuya media depende linealmente de un vector **z** de covariables. Por ejemplo, si $z_i$ es una variable explicativa escalar para la observación i, entonces la media de la distribución normal es $\mu_i = \beta_0 + \beta_1 \cdot z_i$, a menudo pensamos en $z_i$ como una parte fijada (valor determinístico) del i–ésimo dato.

### 5.1.1. Modelo de tiempo de vida acelerada

La mayoría de los mecanismos modernos están diseñados para operar sin fallo durante largos periodos de tiempo, así, únicamente un número reducido de unidades fallará o se deteriorará apreciablemente durante un experimento de fiabilidad en condiciones normales de funcionamiento. Por ejemplo, en el diseño y construcción de un cable de comunicaciones sólo se tienen seis meses para probar componentes que se espera estén operativas durante años. En estas situaciones, los tests (modelos) de vida acelerada resultan adecuados con objeto de obtener información acerca de las características de fiabilidad de este tipo de sistemas, es decir, los ensayos de fiabilidad son *acelerados* en uno u otro sentido.

***Aumentar la tasa de uso del dispositivo*.** Por ejemplo, en un experimento sobre la vida de la tapa de una lavadora de ropa, se podría usar la tapa 100 veces al día. Suponiendo un perfil típico de trabajo, de un ciclo de uso por día, un experimento con una duración de dos semanas podría estimar la distribución de tiempo de fallo correspondiente a 1400 días (casi cuatro años). Este método de aceleración supone que la frecuencia o tasa de uso se puede modelar adecuadamente por ciclos de operación y que no afecta a la distribución del tiempo de fallo.

***Endurecer las condiciones de uso.*** El fallo de multitud de mecanismos está originado por degradación química que puede acelerarse endureciendo las condiciones ambientales con respecto a las habituales en el uso normal del mecanismo. Por ejemplo, el fallo de un determinado mecanismo puede acelerarse probando a temperaturas más altas de las usuales. En este caso hay que suponer que el tiempo de fallo (tipo de distribución) del mecanismo es el mismo a todas las temperaturas a las que se realiza el experimento. En otros casos las condiciones de uso pueden referirse a esfuerzos mecánicos, voltaje, presión, etc.

Todos estos métodos asumen la existencia de un modelo que relaciona la variable aceleradora con la vida útil de la componente, lo cual permite, por medio de extrapolar a condiciones de uso normales, los resultados obtenidos a condiciones aceleradas. La base de esta extrapolación está en un modelo físico/químico para los fallos. Los más usados son el modelo de Arrhenius y la regla de potencia inversa (ver Nelson 1990). Se trata de considerar un factor de aceleración del tiempo de fallo (vida) que resulta ser una función de **z**, de la siguiente forma

$$P\{T_{\mathbf{z}} > t\} = P\{T_0 > t \cdot \psi(\mathbf{z})\},$$

donde estamos denotando $T_{\mathbf{z}}$ al tiempo de funcionamiento bajo las condiciones especificadas en **z**; $T_0$ es el tiempo correspondiente a unas condiciones de funcionamiento que llamaremos "condiciones base", es decir $\mathbf{z} = \mathbf{z_0}$; $\psi(\mathbf{z})$ es lo que llamaremos *función enlace* o, en este caso, *factor de aceleración* y satisface las condiciones

- $\psi(\mathbf{z}) > 0$;
- $\psi(\mathbf{z_0}) = 1$.

La expresión anterior es equivalente, en términos de las correspondientes funciones de supervivencia o fiabilidad de $T_{\mathbf{z}}$ y $T_0$, a

$$\overline{F}_{\mathbf{z}}(t) = \overline{F}_0\left(t \cdot \psi(\mathbf{z})\right),$$

es decir $T_{\mathbf{z}} =_{st} \dfrac{T_0}{\psi(\mathbf{z})}$, son estocásticamente equivalentes. Si $\psi(\mathbf{z}) \neq 1$, las funciones de supervivencia no se cruzan.

A continuación expresamos el modelo en términos de otras funciones de interés, en concreto las razones de fallo se relacionan mediante

$$r_{\mathbf{z}}(t) = r_0\left(t \cdot \psi(\mathbf{z})\right) \cdot \psi(\mathbf{z});$$

y la relación entre los cuantiles viene dada por

$$t_{p,\mathbf{z}} = \frac{t_{p,0}}{\psi(\mathbf{z})},$$

tomando logaritmos

$$\log\left(t_{p,\mathbf{z}}\right) = \log\left(t_{p,0}\right) - \log\left(\psi(\mathbf{z})\right),$$

de modo que en un gráfico de probabilidad con escala logarítmica, $F_{\mathbf{z}}(t)$ es una traslación de $F_0(t)$ a lo largo del eje log(t).

Si $\psi(\mathbf{z}) > 1$, las covariables aceleran el tiempo, en el sentido de que el ítem se mueve más rápidamente a través del tiempo al nivel $\mathbf{z}$ de lo que lo hace en las condiciones base, es decir cuando $\mathbf{z} = \mathbf{z}_0$. Por el contrario, si $\psi(\mathbf{z}) < 1$, las covariables desaceleran el tiempo, en el sentido de que el ítem se mueve más lentamente a través del tiempo al nivel $\mathbf{z}$ de lo que lo hace en las condiciones base, es decir cuando $\mathbf{z} = \mathbf{z}_0$.

Una elección habitual para la función enlace es

$$\psi(\mathbf{z}) = e^{\beta' \mathbf{z}} = \exp\left\{\beta_1 z_1 + \ldots + \beta_k z_k\right\}.$$

Este modelo se usa generalmente en situaciones prácticas en las se pretende obtener rápidamente información acerca de la distribución del tiempo de funcionamiento de determinados ítems. Estas pruebas consisten en someter los ítems a unas condiciones de funcionamiento más severas que las presentes en su uso normal, es decir, se fuerza a los ítems a soportar niveles más altos de temperatura, voltaje, presión, vibración, carga, etc., o combinaciones de éstos. Así, los datos obtenidos a estas condiciones de aceleración son extrapolados por medio de un modelo apropiado a las condiciones normales para obtener una estimación de la distribución del tiempo de vida en esas condiciones normales de funcionamiento. Este procedimiento conlleva un ahorro económico y de tiempo comparado con el experimento realizado en las condiciones reales.

### 5.1.2. Modelo riesgos proporcionales

Este modelo fue introducido por Cox en 1972 y en él se establece la siguiente relación

$$r_{\mathbf{z}}(t) = \psi(\mathbf{z}) r_0(t),$$

para todo $t > 0$; siendo $r_{\mathbf{z}}$ la función de riesgo de un sujeto bajo las condiciones $\mathbf{z}$, y $r_0$ la función de riesgo base. A diferencia del modelo de tiempo de vida acelerada, en este caso el efecto de las covariables es multiplicativo pero sobre la función de riesgo y no

directamente sobre el tiempo. Es decir, ahora las covariables modifican el riesgo de fallo o muerte del ítem con respecto a las condiciones base. También en este caso $\psi(\mathbf{z})$ es una función positiva tal que $\psi(\mathbf{z}_0) = 1$. La relación expresada en términos de las funciones de supervivencia sería en este caso

$$\overline{F}_{\mathbf{z}}(t) = \left[\overline{F}_0(t)\right]^{\psi(\mathbf{z})},$$

de nuevo estas funciones no se cruzan cuando $\psi(\mathbf{z}) \neq 1$.

Si $\psi(\mathbf{z}) > (<) 1$, las covariables incrementan (disminuyen) el riesgo de fallo, y en este caso el modelo acelera (desacelera) el tiempo en el sentido de que $\overline{F}_{\mathbf{z}}(t) > (<) \overline{F}_0(t)$. De nuevo una elección apropiada y habitual para la función enlace es la exponencial, es decir

$$\psi(\mathbf{z}) = e^{\beta' \mathbf{z}} = \exp\{\beta_1 z_1 + ... + \beta_k z_k\}.$$

A modo de resumen presentamos el siguiente cuadro.

| | MODELO AL | MODELO COX |
|---|---|---|
| $\overline{F}_{\mathbf{z}}(t)$ | $\overline{F}_0(t \cdot \psi(\mathbf{z}))$ | $\left[\overline{F}_0(t)\right]^{\psi(\mathbf{z})}$ |
| $f_{\mathbf{z}}(t)$ | $\psi(\mathbf{z}) \cdot f_0(t \cdot \psi(\mathbf{z}))$ | $f_0(t) \cdot \psi(\mathbf{z}) \cdot \left[\overline{F}_0(t)\right]^{\psi(\mathbf{z})-1}$ |
| $r_{\mathbf{z}}(t)$ | $\psi(\mathbf{z}) \cdot r_0(t \cdot \psi(\mathbf{z}))$ | $\psi(\mathbf{z}) \cdot r_0(t)$ |
| $\Lambda_{\mathbf{z}}(t)$ | $\Lambda_0(t \cdot \psi(\mathbf{z}))$ | $\psi(\mathbf{z}) \cdot \Lambda_0(t)$ |

Este modelo recibe el nombre de riesgos proporcionales porque el cociente de las funciones de riesgo de dos individuos con vectores de covariables $\mathbf{z}_1$ y $\mathbf{z}_2$ respectivamente verifica

$$\frac{r_{\mathbf{z}_1}(t)}{r_{\mathbf{z}_2}(t)} = \frac{\psi(\mathbf{z}_1) r_0(t)}{\psi(\mathbf{z}_2) r_0(t)} = \frac{\psi(\mathbf{z}_1)}{\psi(\mathbf{z}_2)} = e^{(\mathbf{z}_1 - \mathbf{z}_2)' \beta}.$$

Es decir, el cociente de funciones de riesgo es constante con el tiempo y por lo tanto, las funciones de riesgo son proporcionales. En el caso particular en que z es un escalar que toma sólo valores 1 y 0, por ejemplo supongamos que se está estudiando eficacia de determinado tratamiento en la evolución de determinada enfermedad y pensemos en z como una variable que toma el valor 1 si el tratamiento es administrado a un paciente y 0 si no lo es. La variable de interés podría ser el tiempo de recaída. En este caso, el cociente anterior se reduce a

$$\frac{r_{\mathbf{z}=1}(t)}{r_{\mathbf{z}=0}(t)} = e^{\beta}.$$

Este cociente indica el riesgo de recaída que un paciente que está recibiendo el tratamiento tiene comparado con un paciente al que no se le administra dicho tratamiento, esta cantidad se denomina riesgo relativo. Si el coeficiente estimado $\beta$ es menor que 0, el riesgo relativo es menor que 1 y por lo tanto estaríamos concluyendo que el uso del tratamiento reduce el riesgo de recaída de la enfermedad en un $(1-e^{\beta})100\%$. En cambio, un coeficiente mayor que 0 nos daría un riesgo relativo mayor que 1, por lo tanto la conclusión sería que con dicho tratamiento se está incrementando el riesgo de recaer en un $(e^{\beta}-1)100\%$.

## *5.2. Procedimientos de estimación*

### 5.2.1. Estimación en el modelo de tiempo de vida acelerada

Exponemos a continuación una serie de pasos adecuados para el análisis de datos procedentes de un test de vida acelerada.

1. Examinar una representación gráfica del tiempo de vida frente a la variable de aceleración.
2. Ajustar distribuciones individualmente a los datos a los distintos niveles observados de la variable de aceleración. Representar gráficamente, en cada grupo, los puntos y las rectas estimadas por máxima verosimilitud. Es aconsejable repetir este proceso usando diferentes modelos paramétricos (Weibull, lognormal, etc).
3. Ajustar a los datos un modelo de regresión con la relación propuesta entre el tiempo de fallo y la variable aceleración (por ejemplo, Arrhenius–lognormal, potencia inversa, etc., ver el texto de Meeker y Escobar, 1998). En el modelo más general se establece que

   $$T = e^{z'\beta} T_0$$

   o, equivalentemente

   $$\ln T = z'\beta + \ln T_0 ,$$

   donde **z** es el vector de covariables, **β** es el vector de parámetros a estimar y $T_0$ es el tiempo de vida de un ítem en condiciones base, que tendrá una distribución de probabilidad fijada de antemano. Lo más habitual en supervivencia y fiabilidad es asumir una distribución lognormal o de Weibull.

4. Comparar el modelo combinado del paso 3 con los análisis individuales en el paso 2 con el fin de detectar falta de ajuste.
5. Desarrollar análisis de residuos y otras comprobaciones sobre las hipótesis del modelo.
6. Test para comprobar la validez del modelo. Se trata de comparar los análisis individuales con el modelo. Se considera un test de razón de verosimilitudes que proporciona una valoración analítica sobre si las desviaciones observadas entre el ajuste del modelo individual (modelo sin restricciones) y el modelo considerando la variable de aceleración (modelo restringido) puede deberse al azar o no. Llamamos $\Lambda_{total}$, a la suma de las log–verosimilitudes obtenidas considerando por separado cada grupo determinado por los diferentes niveles de estrés (modelo sin restricciones); y, $\Lambda_{alt}$, a la log–verosimilitud en el modelo restringido. Si el modelo de regresión es "correcto", el estadístico

   $$Q = -2\left(L_{alt} - L_{total}\right)$$

   tiene distribución $\chi^2(m-k)$, siendo m el número total de parámetros estimados en el modelo sin restricciones y k el número de parámetros estimados en el modelo de regresión.

### 5.2.2. Estimación en el modelo de riesgos proporcionales

Ilustramos este caso suponiendo el modelo más sencillo, en el que la función de riesgo base es constante, es decir, corresponde a una distribución exponencial. Si disponemos de una muestra censurada de tamaño n, en la que los elementos señalados con el

subíndice $i \in D$ son observaciones completas, y los de subíndice $i \in C$ son datos censurados, podemos construir la función de verosimilitud obteniendo

$$L(\boldsymbol{\beta}) = \left( \prod_{i \in D} e^{\mathbf{z}_i \boldsymbol{\beta}} e^{-t_i \exp(z_i \boldsymbol{\beta})} \right) \times \left( \prod_{i \in C} e^{-t_i \exp(z_i \boldsymbol{\beta})} \right),$$

siendo $\boldsymbol{\beta} = (\beta_{i0}, \beta_{i1}, \ldots, \beta_{im})'$ el vector de parámetros de regresión y $\mathbf{z}_i = (z_{i0}, z_{i1}, \ldots, z_{im})$ el vector de covariables que describe al individuo i–ésimo. Por lo tanto, tomando logaritmos

$$\Lambda(\boldsymbol{\beta}) = \sum_{i \in D} \mathbf{z}_i \boldsymbol{\beta} - \sum_{i=1}^{n} t_i \exp(\mathbf{z}_i \boldsymbol{\beta}).$$

Las ecuaciones de verosimilitud vienen dadas por

$$\frac{\partial \Lambda(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i \in D} z_{ij} - \sum_{i=1}^{n} t_i z_{ij} \exp(\mathbf{z}_i \boldsymbol{\beta}) = 0, \qquad j = 0,1,\ldots,m$$

La matriz de información de Fisher observada es

$$I_0(\hat{\boldsymbol{\beta}}) = \left[ -\frac{\partial^2 \Lambda(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} \right]_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}} = \left[ \sum_{i=1}^{n} t_i z_{ij} z_{ik} \exp(\mathbf{z}_i \boldsymbol{\beta}) \right]_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}}, \qquad j, k = 0, 1, 2, \ldots, m.$$

Para resolver las ecuaciones de verosimilitud usaremos un algoritmo basado en el método de Newton–Raphson:

1. Sea $\boldsymbol{\beta}^0$ un valor del parámetro;
2. Sea $\mathbf{U}(\boldsymbol{\beta})$ el vector cuyas componentes son $U_j(\boldsymbol{\beta}) = \dfrac{\partial \Lambda(\boldsymbol{\beta})}{\partial \beta_j}$, para j = 0, 1,2,…,m;
3. Sea $F(\boldsymbol{\beta})$ la matriz formada por $F_{jk}(\boldsymbol{\beta}) = \dfrac{\partial^2 \Lambda(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k}$, con j, k = 0, 1,2,…,m;
4. El desarrollo en serie de Taylor hasta el segundo orden del vector $\mathbf{U}(\boldsymbol{\beta})$ alrededor de $\boldsymbol{\beta}^0$ viene dada por
$$\mathbf{U}(\boldsymbol{\beta}) = \mathbf{U}(\boldsymbol{\beta}^0) + \mathbf{F}(\boldsymbol{\beta}^0) \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}^0);$$
5. Si sustituimos el valor de $\hat{\boldsymbol{\beta}}$ y lo despejamos en la expresión anterior, tendríamos, ya que $U(\hat{\boldsymbol{\beta}}) = 0$,
$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^0 + \left[ I(\boldsymbol{\beta}^0) \right]^{-1} \mathbf{U}(\boldsymbol{\beta}^0)$$
6. Usamos el siguiente método iterativo (Newton–Raphson)
$$\boldsymbol{\beta}^{(n+1)} = \boldsymbol{\beta}^{(n)} + \left[ I(\boldsymbol{\beta}^{(n)}) \right]^{-1} \mathbf{U}(\boldsymbol{\beta}^{(n)}), \quad n \geq 0.$$
Debemos establecer un criterio de parada en función de que $\boldsymbol{\beta}^{(n+1)} \approx \boldsymbol{\beta}^{(n)}$, y también elegir un valor para la semilla $\boldsymbol{\beta}^0$, que sea adecuado.