

Capítulo 4

El uso de información concomitante

4.1. Datos de supervivencia con variables explicativas

En muchas situaciones prácticas, los datos de supervivencia vienen con información asociada o *concomitante* de la cual se cree que dependen los tiempos de supervivencia. En datos de medicina, por ejemplo, esta información puede ser por diseño, en el sentido de que uno o mas tratamientos se comparan con el efecto de tiempos de supervivencia de un grupo de “control”; o puede ocurrir al clasificar pacientes de alguna forma natural, como es género, grupo de edad, etc.

En la mayoría de las situaciones en que se observan tiempos de supervivencia habrá alguna, o a veces mucha, información auxiliar. Esta información puede venir en la forma de una observación de una variable continua, o bien en la forma de una variable categórica. Para entender cómo incluir esta información en los modelos de supervivencia, a continuación se da un breve resumen de su uso en el contexto de los modelos lineales generalizados.

4.2. El Modelo lineal generalizado

Supóngase que se tiene una variable respuesta Y , que es la variable aleatoria de interés, y varias variables explicativas. El conocimiento del contexto en el cual se obtuvieron los datos - tales como relaciones teóricas entre las variables, el diseño del estudio y los resultados de un análisis exploratorio de los datos - pueden hacerse para formular el modelo. El modelo lineal tiene dos componentes:

1. La función de distribución de Y , e.g. $Y \sim N(\mu, \sigma^2)$, $Y \sim \text{EXP}(\theta)$.
2. Una ecuación que *liga* el valor de Y con una combinación lineal de las variables explicativas, e.g. $E[Y] = \beta_0 + \beta_1 x$ o $\log(E(Y)) = \beta_0 + \beta_1 \text{sen}(\alpha x)$.

Los modelos lineales generalizados son aquellos cuyas funciones de densidad pertenecen a la familia exponencial de distribuciones, las cuales incluyen la normal, binomial, Poisson etc. La ecuación en el segundo componente tiene la forma general:

$$g[E(Y)] = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

donde la parte $\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$ se llama el componente lineal.

Para respuestas Y_1, \dots, Y_n , i.e. para una muestra aleatoria, esta forma general se puede escribir en notación de matrices como

$$g(E[Y]) = \mathbf{X}\boldsymbol{\beta}$$

donde

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

es el vector de respuestas,

$$g(\mathbf{E}[\mathbf{Y}]) = \begin{pmatrix} g[\mathbf{E}(Y_1)] \\ \vdots \\ g[\mathbf{E}(Y_n)] \end{pmatrix}$$

es el vector de funciones de los términos $E[Y_i]$ (con la misma g para cada elemento),

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

es el vector de parámetros, y \mathbf{X} es la matriz cuyos elementos son constantes que representan niveles de variables explicativas categóricas o variables explicativas de medidas continuas.

Para una variable explicativa continua x (tal como edad) el modelo contiene un término $\beta_1 x$ donde el parámetro β representa el cambio de la respuesta correspondiente al cambio de una unidad en x .

Para una variable explicativa categórica hay parámetros para los diferentes niveles de un *factor*. Los elementos correspondientes en \mathbf{X} se eligen para excluir o incluir los parámetros apropiados de cada observación. Estas variables son llamadas *variables ficticias* o *variables indicadoras* (en inglés *dummy variable*); si se componen de ceros y unos, se usa el término *variable indicador*.

Si hay $(p + 1)$ parámetros en el modelo y N observaciones, entonces Y es un vector aleatorio de $n + 1$, $\boldsymbol{\beta}$ es un vector de $p + 1$ parámetros y \mathbf{X} es una matriz de constantes conocidas. A \mathbf{X} se le llama la *matriz diseño* y $\mathbf{X}\boldsymbol{\beta}$ es conocido como el componente lineal.

Ejemplo 4.1. Considérese el modelo de regresión lineal simple para dos grupos (e.g. por sexo) con el mismo número de observaciones en cada grupo

$$g(\mathbf{E}[Y_{jk}]) = g(\mu_{jk}) = \alpha_j + \beta_j X_{jk},$$

donde $j = 1, 2$ y $k = 1, \dots, K$. Aquí el vector de respuestas y el vector de coeficientes toman la forma

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1K} \\ Y_{21} \\ \vdots \\ Y_{2K} \end{pmatrix}, \quad \text{y} \quad \boldsymbol{\beta} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{pmatrix},$$

y la matriz diseño correspondiente es

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & X_{11} & 0 \\ 1 & 0 & X_{12} & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & X_{1K} & 0 \\ 0 & 1 & 0 & X_{21} \\ 0 & 1 & 0 & X_{22} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & X_{2K} \end{pmatrix}.$$

Ejemplo 4.2. Considérense formulaciones alternativas para comparar los efectos de dos grupos. La muestra aleatoria tiene la forma $Y_{11}, \dots, Y_{1K_1}, Y_{21}, \dots, Y_{2K_2}$.

a) Si $g(E[Y_{1j}]) = \beta_1$ y $g(E[Y_{2k}]) = \beta_2$, entonces

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1K_1} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2K_2} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \quad \text{y} \quad \mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix}.$$

b) Si $g(E[Y_{1j}]) = \mu + \alpha_1$ y $g(E[Y_{2k}]) = \mu + \alpha_2$, μ representa la media general y α_1

y α_2 son las diferencias a partir de μ . De esta forma:

$$\beta = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix} \quad y \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{pmatrix}.$$

Esta formulación, sin embargo, no se recomienda. El problema es que se tienen muchos parámetros dentro del modelo. Es, entonces, necesario hacer una modificación.

- c) Si $g(E[Y_{1K}]) = \mu$ y $g(E[Y_{2K}]) = \mu + \alpha$, el grupo 1 se trata como el grupo de referencia y α representa el efecto adicional del grupo 2. Entonces

$$\beta = \begin{pmatrix} \mu \\ \alpha \end{pmatrix} \quad y \quad \mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix}$$

- d) Si $g(E[Y_{1K}]) = \mu + \alpha$ y $g(E[Y_{2K}]) = \mu - \alpha$, entonces los grupos se tratan simétricamente; μ es el efecto promedio general y α representa las diferencias del grupo. Aquí,

$$\beta = \begin{pmatrix} \mu \\ \alpha \end{pmatrix} \quad y \quad \mathbf{X} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & -1 \\ \vdots & \vdots \\ 1 & -1 \end{pmatrix}$$

Ejemplo 4.3. Variables explicativas ordinales. Supóngase que los datos se obtienen para tres grupos de pacientes con una enfermedad poco ayuda, moderada o severa. Los grupos pueden describirse por niveles de una variable ordinal. Esta se

puede describir como:

$$\begin{aligned}
 g(\mathbb{E}[Y_{1j}]) &= \mu, \\
 g(\mathbb{E}[Y_{2k}]) &= \mu + \alpha, \\
 g(\mathbb{E}[Y_{3l}]) &= \mu + \alpha_1 + \alpha_2,
 \end{aligned}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix}, \quad \text{y} \quad \mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 1 \end{pmatrix}$$

Así, α_1 representa el efecto del grupo 2 relativo al grupo 1 y α_2 representa el efecto del grupo 3 relativo al grupo 2.

Inclusión de un factor

Un factor es, entonces, una variable explicativa asociada con una variable categórica con J niveles o formas de agrupar las observaciones. Un método conveniente para incluir un factor en un modelo es considerar los efectos principales del factor $\alpha_1, \alpha_2, \dots, \alpha_J$, de manera tal que

$$g(\mathbb{E}[Y_{jk}]) = g(\mu_j) = \mu + \alpha_j, \quad j = 1, \dots, J,$$

con la restricción $\alpha_1 = 0$.

Bajo esta estructura, los términos $\mu, \alpha_2, \dots, \alpha_J$ pueden modelarse al definir J variables indicadoras $X_0, X_2, X_3, \dots, X_J$ para obtener el modelo lineal

$$\mathbb{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}.$$

La matriz diseño puede entonces construirse de acuerdo a la siguiente descripción dependiendo del nivel a que pertenezca cada observación:

Nivel de A	X_0	X_2	X_3	\cdots	X_J
1	1	0	0		0
2	1	1	0	\cdots	0
3	1	0	1	\cdots	\vdots
\vdots	\vdots			\ddots	\vdots
J	1	0	\cdots		1

De esta forma, el modelo queda como

$$g(\mathbb{E}[Y]) = \mu x_0 + \alpha_2 x_2 + \cdots + \alpha_J x_J$$

donde x_j es el valor de X_j para un individuo que en particular. Cuando se tiene la restricción $\alpha_1 = 0$, se tienen contrastes de tratamientos, en el lenguaje S-PLUS, es posible invocar este formato con el comando:

```
options(contrasts = c("contr.treatment", "contr.poly"))
```

El lenguaje R tiene esta estructura siempre.

Inclusión de dos factores

Considérense el factor A con J niveles y el factor B con K niveles. Es posible clasificar los datos en forma *cruzada* con los JK subgrupos que forman todas las combinaciones de los niveles de A y B.

Aquí las respuestas son del estilo Y_{jkl} , la l -ésima observación del k -ésimo nivel del factor B del j -ésimo nivel del factor A. Las siguientes estructuras forman todas las posibilidades de incluir los factores en el modelo:

i) El *modelo nulo*:

$$g(\mathbb{E}[Y_{jkl}]) = \mu,$$

que es el modelo más simple.

ii) El modelo A:

$$g(\mathbb{E}[Y_{jkl}]) = \mu_j, \quad j = 1, \dots, J,$$

o en forma equivalente

$$g(\mathbb{E}[Y_{jkl}]) = \mu + \alpha_j,$$

con alguna de las siguientes restricciones

a) $\alpha_1 = 0$, ó

b) $\sum \alpha_j = 0$.

iii) El modelo B:

$$g(\mathbb{E}[Y_{jkl}]) = \mu + \beta_k, \quad k = 1, \dots, K,$$

con alguna de las siguientes restricciones

a) $\beta_1 = 0$ ó

b) $\sum \beta_k = 0$

iv) El modelo $A + B$:

$$g(\mathbb{E}[Y_{jkl}]) = \mu + \alpha_j + \beta_k \quad j = 1, \dots, J, \quad k = 1, \dots, K,$$

donde la restricción es

i) $\alpha_1 = \beta_1 = 0$ ó

ii) $\sum \alpha_j = \sum \beta_k = 0$.

v) El *modelo saturado* $A * B$ con *interacción*:

$$g(\mathbb{E}[Y_{jkl}]) = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} \quad j = 1, \dots, J, \quad k = 1, \dots, K;$$

Tabla 4.1: Modelos anidados para los factores A y B.

Fórmula	Modelo	número de parámetros en β
$A * B$	$\mu + \alpha_j + \beta_k + (\alpha\beta)_{jk}$	JK
$A + B$	$\mu + \alpha_j + \beta_k$	$JK - (J - 1)(K - 1)$
A	$\mu + \alpha_j$	J
B	$\mu + \beta_k$	K
nulo	μ	1

aquí, $(\alpha\beta)_{jk}$ corresponde a los efectos de interacción, y α_j y β_k corresponden a los efectos principales de A y B respectivamente.

La Tabla 4.1 muestra un resumen de los modelos generados con los factores A y B. Es posible observar que los modelos están relacionados con ciertos parámetros, lo cual es útil para llevar a cabo pruebas de significancia.

Las interacciones son términos en el modelo que corresponden a efectos individuales para cada combinación de niveles de los factores. Para incluir el término $(\alpha\beta)_{jk}$ en el modelo, se calculan los productos de las variables indicadoras incluidos en los efectos principales. De esta forma, existen $(J - 1)(K - 1)$ parámetros asociados con la interacción. Si, por ejemplo, $J = 2$ y $K = 2$, entonces el vector de coeficientes del modelo saturado es

$$\beta = \begin{pmatrix} \mu \\ \alpha_2 \\ \beta_2 \\ (\alpha\beta)_{22} \end{pmatrix}.$$

Inclusión de un factor y una variable continua

Hay situaciones en las que se puede estar interesado en comparar medias de subgrupos definidos por los niveles de un factor pero reconociendo que variables

contínuas pueden afectar las respuestas. En este caso, $g(E[Y_{jk}])$ se puede modelar como una línea recta de una variable continua x_{jk} con la posibilidad de asignar pendientes y ordenadas al origen a cada subgrupo del factor A.

A continuación se describen los modelos que se pueden formar con una factor A y una variable continua c.

i) El modelo nulo:

$$g(E[Y_{jk}]) = \mu,$$

el modelo más simple.

ii) El modelo A:

$$E[Y_{jk}] = \mu + \alpha_j \quad j = 1, \dots, J,$$

con la restricción $\alpha = 0$.

iii) El modelo c:

$$g(E[Y_{jk}]) = \mu + \gamma x_{jk},$$

el ajuste de una recta con la variable explicativa x_{jk} para $g(E[Y_{jk}])$.

iv) El modelo A+c:

$$E[Y_{jk}] = \mu + \alpha_j + \gamma X_{jk},$$

donde $\alpha_1 = 0$, el modelo con diferentes ordenadas al origen y pendiente común para todos los niveles del factor.

v) El modelo A*c:

$$E[Y_{jk}] = \mu + \alpha_j + \gamma X_{jk} + (\alpha\gamma)_j X_{jk},$$

donde $\alpha_1 = 0$, el modelo que considera diferentes pendientes y ordenadas al origen para cada subgrupo en A. Aquí, $(\alpha\gamma)_j$ representa el coeficiente correspondiente a la interacción entre A y c.

En general, es posible incluir los factores y variables continuas que se deseen, siempre y cuando el número de coeficientes en β no supere el tamaño de la muestra. La interpretación de los modelos son generalizaciones de los resultados aquí mostrados. Cuando se trate de pronosticar $g(E[Y])$ para un individuo en particular, sólo se necesita sustituir los valores correspondientes en cada una de las variables explicativas del modelo estimado.

4.3. Inferencia para Modelos de regresión paramétrica

4.3.1. Las distribuciones Exponencial y Weibull

Para modelar una distribución exponencial, la dependencia del tiempo de supervivencia T , $T \sim \text{EXP}(\theta)$, de las variables explicativas se puede modelar con

$$\log(E[T]) = \log(\theta) = \mathbf{x}^T \boldsymbol{\beta},$$

donde $\mathbf{x}^T = (1, x_1, \dots, x_p)$ es el vector de variables explicativas y $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \dots, \beta_p)$ es el vector de coeficientes correspondiente. En este caso, la función liga g es el logaritmo natural. Esta liga es conveniente pues asegura que el parámetro θ permanezca siempre positivo. Nótese que β_0 es el coeficiente correspondiente a la ordenada al origen.

La distribución Weibull de dos parámetros, a diferencia de la Exponencial, no pertenece a la familia de distribuciones exponencial; por lo cual, no es posible aplicar una liga directamente a la media de los tiempos de supervivencia como se ha descrito en la sección anterior. Sin embargo, es posible tomar algunas ideas de los modelos lineales generalizados y entonces incluir un componente lineal en alguno de los parámetros asociados con la distribución.

Se ha visto que una de las parametrizaciones de la función de supervivencia del

modelo Weibull es:

$$S_T(t) = \exp(-\lambda t^\alpha),$$

y la fuerza de mortalidad correspondiente se expresa como

$$\hat{h}_T(t) = \lambda \alpha t^{\alpha-1}.$$

Si se toma la transformación $Y = \log T$, la función de supervivencia correspondiente de Y es

$$S_Y(y) = \exp(-\lambda e^{\alpha y}).$$

Si se redefinen los parámetros con $\lambda = \exp\{-\mu/\sigma\}$ y $\sigma = 1/\alpha$, entonces Y toma la forma de un modelo log lineal con

$$Y = \log T = \mu + \sigma W,$$

donde W es la distribución de valor extremo con función de densidad dada por

$$f_W(w) = \exp\{w - e^w\}$$

y función de supervivencia,

$$S_W(w) = \exp(-e^w).$$

De esta forma, las funciones de densidad y de supervivencia de Y están dadas por

$$f_Y(y) = (1/\sigma) \exp[(y - \mu)/\sigma - e^{(y-\mu)/\sigma}]$$

y

$$S_Y(y) = \exp(-e^{(y-\mu)/\sigma})$$

respectivamente. Bajo esta estructura, se modelan tiempos de supervivencia T , donde $T \sim \text{WEI}(\exp\{\mathbf{x}^T \boldsymbol{\beta}\}, 1/\sigma)$. Cuando $\alpha = 1$, o en forma equivalente, cuando $\sigma = 1$, entonces la distribución Weibull se reduce a la distribución exponencial.

La función de verosimilitud para datos con censura por la derecha está dada por

$$\begin{aligned} L &= \prod_{j=1}^n [f_Y(y_j)]^{\delta_j} [S_Y(y_j)]^{(1-\delta_j)} \\ &= \prod_{j=1}^n \left[f_W \left(\frac{y_j - \mu}{\sigma} \right) \right]^{\delta_j} \left[S_W \left(\frac{y_j - \mu}{\sigma} \right) \right]^{(1-\delta_j)}. \end{aligned}$$

Una vez que se han encontrado los estimadores de máxima verosimilitud para los parámetros μ y σ , entonces los de λ y α , se pueden obtener estimadores para las funciones de supervivencia y de fuerza de mortalidad para T o Y .

Los estimadores de μ y α se encuentran numéricamente con rutinas ya disponibles. La matriz de covarianzas de los parámetros μ y α están disponibles en los paquetes estadísticos también. Usando la propiedad de invarianza de los estimadores de máxima verosimilitud, se tiene que los estimadores de máxima verosimilitud de λ y α son

$$\hat{\lambda} = \exp(-\hat{\mu}/\hat{\sigma}) \quad \text{y} \quad \hat{\alpha} = 1/\hat{\sigma}.$$

Para obtener las varianzas y covarianzas de $\hat{\lambda}$ y $\hat{\alpha}$ es posible usar las aproximaciones de series de Taylor; estas son:

$$\text{Var}[g(\hat{\theta}_1, \hat{\theta}_2)] \approx \left(\frac{\partial g}{\partial \hat{\theta}_1} \right)^2 \text{Var}[\hat{\theta}_1] + \left(\frac{\partial g}{\partial \hat{\theta}_2} \right)^2 \text{Var}[\hat{\theta}_2] + 2 \left(\frac{\partial g}{\partial \hat{\theta}_1} \frac{\partial g}{\partial \hat{\theta}_2} \right) \text{Cov}[\hat{\theta}_1, \hat{\theta}_2].$$

y

$$\begin{aligned} \text{Cov}[g_1(\hat{\theta}_1, \hat{\theta}_2), g_2(\hat{\theta}_1, \hat{\theta}_2)] &\approx \left(\frac{\partial g_1}{\partial \hat{\theta}_1} \frac{\partial g_2}{\partial \hat{\theta}_1} \right) \text{Var}[\hat{\theta}_1] + \left(\frac{\partial g_1}{\partial \hat{\theta}_2} \frac{\partial g_2}{\partial \hat{\theta}_2} \right) \text{Var}[\hat{\theta}_2] + \\ &\quad \left(\frac{\partial g_1}{\partial \hat{\theta}_1} \frac{\partial g_2}{\partial \hat{\theta}_2} + \frac{\partial g_1}{\partial \hat{\theta}_2} \frac{\partial g_2}{\partial \hat{\theta}_1} \right) \text{Cov}[\hat{\theta}_1, \hat{\theta}_2]. \end{aligned}$$

De esta forma,

$$\text{Var}(\hat{\lambda}) = \exp\{-2\hat{\mu}/\hat{\sigma}\} [\text{Var}(\hat{\mu})/\hat{\sigma}^2 + \hat{\mu}^2 \text{Var}(\hat{\sigma})/\hat{\sigma}^4 - 2\hat{\mu} \text{Cov}(\hat{\mu}, \hat{\sigma})/\hat{\sigma}^3],$$

$$\text{Var}(\hat{\alpha}) = \text{Var}(\hat{\sigma})/\hat{\sigma}^4,$$

Tabla 4.2: Estimadores del modelo log lineal Weibull con el factor estado y la variable edad para los pacientes de cáncer de la laringe

Variable	Estimador del Parámetro	Error Estándar	Ji cuadrada de Wald	Valor p
Ordenada $\hat{\beta}_0$	3.53	0.90		
Escala $\hat{\sigma}$	0.88	0.11		
Z_1 : Etapa II ($\hat{\beta}_1$)	-0.15	0.41	0.13	0.717
Z_2 : Etapa III ($\hat{\beta}_2$)	-0.59	0.32	3.36	0.067
Z_3 : Etapa IV ($\hat{\beta}_3$)	-1.54	0.36	18.07	< 0.0001
Z_4 : Edad ($\hat{\beta}_4$)	-0.02	0.01	1.87	0.172

y

$$\text{Cov}(\hat{\lambda}, \hat{\alpha}) = \exp(-\hat{\mu}/\hat{\sigma})[\text{Cov}(\hat{\mu}, \hat{\sigma})/\hat{\sigma}^3 - \hat{\mu}\text{Var}(\hat{\sigma})/\hat{\sigma}^4].$$

Para incorporar variables auxiliares al modelo Weibull, es posible usar el formato lineal del logaritmo natural del tiempo, de manera tal que:

$$Y = \mathbf{x}^T \boldsymbol{\beta} + \sigma W,$$

donde $\mathbf{x}^T = (1, x_1, \dots, x_p)$ es el vector de variables explicativas y $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \dots, \beta_p)$ es el vector de coeficientes.

Ejemplo 4.4. Considere el estudio de 90 personas del sexo masculino con diagnóstico de cáncer en la laringe (Kardaun, 1983). Se registraron los intervalos (en años) entre el primer tratamiento y la muerte o el término del estudio. También se registró la edad de cada individuo en el momento del diagnóstico, y se usó un criterio médico para clasificar cuatro etapas del cáncer. Los datos se encuentran en el archivo `larynx.dat`. Se desea ajustar el modelo Weibull con la especificación

$$Y = \log T = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \sigma W,$$

donde x_i , $i = 1, \dots, 3$ son los indicadores de las etapas II, II y IV del cáncer, y x_4 es la edad del paciente. Los estimadores de los parámetros, los errores estándares, los estadísticos ji-cuadrada de Wald, y los valores p para probar que $\beta_i = 0$ están dados en la Tabla 4.2. A continuación se muestra la rutina en el lenguaje R para obtener la regresión de los datos de cáncer de laringe:

```
# Invoca la libreria de supervivencia
library(survival)

# Pon en un objeto los datos
larynx <- read.table("C:/Mis documentos/larynx.dat",
                    col.names=c("estado", "tiempo", "edad", "fecha", "estatus"))
# Crea el factor ETAPA en los datos larynx:
larynx$ETAPA <- factor(larynx$estado, labels = c("I", "II", "III", "IV"))

# Ajusta el modelo Weibull:
larynx.regwei <- survreg(Surv(tiempo, estatus)~ ETAPA + edad,
                        data=larynx, dist='weibull')

#Los coeficientes y el estimador de log(sigma) son:
coeficientes <- c(larynx.regwei$coefficients, log(larynx.regwei$scale))

# los errores estandares son:
errores <- sqrt(diag(larynx.regwei$var))

# el estadistico Wald de Ji cuadrada es:
ji.wald <- (coeficientes/errores)^2

# obten el valor p para cada parametro:
valor.p <- 1 - pchisq(ji.wald, 1)
```