Verificación de los modelos de supervivencia de Cox

Preparado por Luis M. Molinero (Alce Ingeniería) Agosto 2004
CorreoE: **bioestadistica@alceingenieria.net**

Artículo en formato PDF
www.seh−lelha.org/stat1.htm

Introducción

El *análisis de supervivencia* nos permite estudiar y construir modelos para analizar el tiempo que un suceso tarda en ocurrir, en los que diferentes variables pronóstico permiten estimar el tiempo de aparición del suceso. Entre los diferentes tipos de modelos que se pueden emplear, uno de los más extendidos en medicina es el **modelo de riesgos proporcionales**, también conocido como modelo de Cox, y del que podemos encontrar publicados gran número de estudios, sobre todo en áreas relativas a enfermedades crónicas (en cardiología es bien conocido el *modelo de Framingham*, y más recientemente el *modelo SCORE*), en trasplante, en oncología, etc.

La información que se puede derivar de estos modelos es:

- ¿Qué variables tienen relevancia como factores pronóstico?
- ¿Cómo podemos clasificar a los pacientes en cuanto a un buen o mal pronóstico?
- ¿Cuál es el pronóstico más probable para un paciente con unas características concretas?

En este documento no vamos a comentar las bases del análisis de supervivencia, ya que están descritas en un artículo anterior, ni tampoco el fundamento de los modelos de Cox que también fueron ya explicados anteriormente. Aquí vamos a partir del punto en el que ya hemos construido dicho modelo de regresión para el tiempo de supervivencia y lo que ahora deseamos es verificar si se cumplen las hipótesis en las que se basa dicho modelo, y si verdaderamente éste se ajusta bien a nuestros datos, dado que se trata de un paso obligado en el proceso de elaboración de un modelo de regresión, y en el caso de los modelos de supervivencia, quizás por desconocimiento o quizás porque es más complicado que en los modelos de regresión lineal, se deja de lado con excesiva frecuencia.

Al igual que en los modelos de regresión lineal, también en los de análisis de supervivencia la mayor parte de los procedimientos de verificación del modelo se basan en unas cantidades denominadas **residuos**. Se denomina residuo a la diferencia entre el valor observado y el valor estimado por la ecuación de regresión, es decir a lo que la ecuación de regresión deja sin explicar para cada paciente. Algunos de los procedimientos para determinar si el modelo está bien construido se basan en representar gráficamente estos residuos y evaluar si presentan patrones anómalos frente a la forma que teóricamente deberían presentar. Por las características matemáticas de los modelos de supervivencia de Cox, el cálculo de los residuos resulta bastante más complicado que en un modelo de regresión lineal, y su deducción matemática es bastante compleja, existiendo diferentes alternativas, así que en este artículo pasaremos de puntillas sobre los detalles matemáticos y únicamente nos centraremos en sus características funcionales, de tal manera que el usuario de los programas de análisis estadístico sepa cómo interpretar los resultados proporcionados por el ordenador, y sepa pedir esos resultados, remitiendo al lector más avezado a la literatura especializada y en su caso a los manuales correspondientes del programa.

Recordemos que en el modelo de regresión de Cox la función de riesgo (*hazard*) se construye como

$$h_i(t) = h_0(t)\exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \qquad [1]$$

Este modelo se denomina *semiparamétrico*, ya que la función de riesgo base o de referencia $h_0(t)$ no queda especificada y puede tomar cualquier forma. Al término $\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$ se le denomina puntuación de riesgo (*risk score*), ya que un valor negativo grande corresponde a un perfil de riesgo menor que la media, mientras que un valor positivo grande de esa puntuación corresponde a un perfil de riesgo mayor que la media. Si tenemos dos perfiles de riesgo diferentes:

$$PR_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

$$PR_k = \beta_1 x_{k1} + \beta_2 x_{k2} + \ldots + \beta_p x_{kp}$$

el cociente de riesgo de un perfil frente a otro (similar al riesgo relativo) es

$$\frac{h_i(t)}{h_k(t)} = \frac{h_0(t)\exp(PR_i)}{h_0(t)\exp(PR_k)} = \frac{\exp(PR_i)}{\exp(PR_k)} \qquad [2]$$

luego el cociente de riesgo entre dos perfiles no depende del tiempo, y por ello a los modelos de Cox se les denomina modelos de riesgo proporcional.

Análisis de residuos

Recordemos que el residuo es una cantidad que se calcula para cada paciente y nos proporciona información en cuanto a la diferencia entre el valor de supervivencia observado para ese paciente y el valor estimado por la ecuación de regresión, cuanto mayor es esa diferencia mayor será el valor del residuo, con su signo correspondiente.

Probablemente en una salida de ordenador de un análisis de supervivencia nos encontraremos con un tipo de residuos denominados **residuos martingala**. Estos se construyen basándose a su vez en los denominados **residuos de Cox−Snell**:

$rm_i = d_i - rc_i$

d_i=1 suceso (muerto)

d_i=0 observación censurada o incompleta

rc_i residuo Cox−Snell

No vamos a entrar en el cálculo matemático de estos residuos, pero sí vamos a comentar las propiedades de los **residuos martingala**. Evidentemente por definición el residuo martingala para un paciente incompleto (censurado) será negativo.

Para los pacientes fallecidos (observaciones completas) el valor de los residuos puede ir desde $-\infty$ hasta 1. Si la muestra es grande la suma de estos residuos es cero, no están correlacionados y el valor esperado es cero. Sin embargo no se distribuyen de forma simétrica en torno a cero, aunque el modelo sea correcto, lo que complica la interpretación de los gráficos.

Por ello se ha definido otro tipo de residuos denominados **residuos de desviación** (deviance). Recordemos que la desviación (deviance) de un modelo de regresión es el estadístico que se utiliza para cuantificar hasta qué punto el modelo actual que hemos estimado se aleja (desvía) de un modelo teórico que se ajustase perfectamente a nuestros datos (denominado *modelo completo* o *modelo saturado*). Cuanto menos se aleje nuestro modelo de ese otro modelo *"ideal"* mejor será el ajuste. La desviación de un modelo se calcula como

$$D = -2\,(\ln L_A - \ln L_S)$$

donde L_A corresponde a las funciones de verosimilitud para el modelo actual y L_S para el modelo saturado. Precisamente la suma de los cuadrados de los residuos de desviación corresponde al valor de la desviación del

modelo $D = \sum rd_i^2$.

Este tipo de residuos de desviación se construyen transformando los residuos martingala de tal manera que produzcan valores simétricos en torno de 0, y ahora el rango de valores va desde $-\infty$ hasta $+\infty$. Sin embargo aunque los residuos de desviación se distribuyen simétricamente en torno de cero si el modelo es adecuado, sin embargo no tienen por qué sumar cero.

Veamos ahora cómo interpretamos los valores de éstos residuos de desviación. Un residuo con un valor negativo grande corresponderá a pacientes que tienen un tiempo de supervivencia grande, y para los que sin embargo el valor estimado por el modelo a partir de los factores pronóstico indica una supervivencia mucho menor. Por el contrario, un residuo con un valor negativo alto corresponde a pacientes con un tiempo de supervivencia pequeño, contrariamente a lo que nos sugiere el modelo.

Si en nuestra base de datos calculamos los residuos para cada paciente y los ordenamos, puede ser interesante revisar los pacientes que tienen valores extremos tanto positivos como negativos ya que, aunque pueden ser correctos, en ocasiones nos permiten detectar errores en la introducción de datos.

Conviene representar los residuos de desviación en el eje Y frente al *índice de paciente* en el eje X (denominamos *índice del paciente* simplemente al número de orden en el que se ha ido registrando cada observación en el estudio o en la base de datos). Obtendremos una gráfica como la de la figura

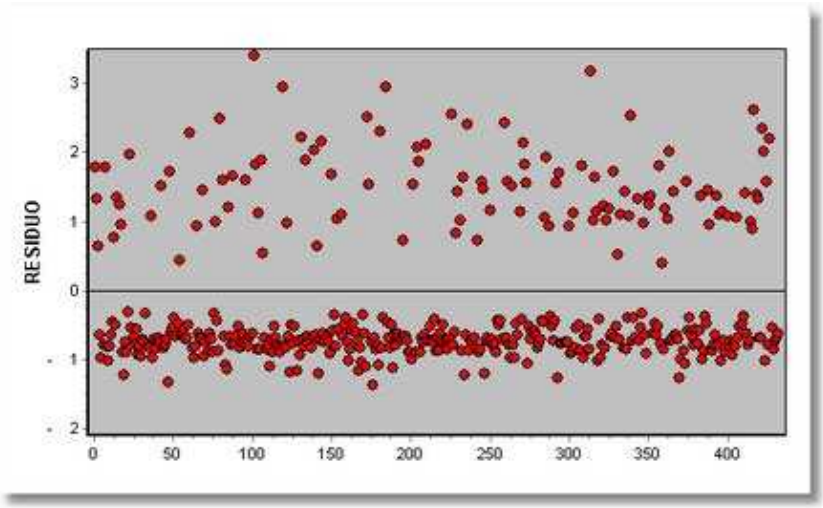

Fig. 1 Residuos de desviación frente al índice de paciente

Aunque no se aprecia ninguna anomalía evidente en la gráfica de residuos de la figura 1, sí vemos que hay algunas observaciones con residuos positivos grandes (en la zona superior) que conviene revisar.

Otro gráfico muy interesante para detectar anomalías en el ajuste consiste en representar en el eje Y el valor de los residuos y en el eje de las X el resultado de calcular la puntuación de riesgo para el modelo estimado (el término $\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$, donde cada β se substituye por el coeficiente estimado y cada X por el valor de esa variable para cada paciente).

En la siguiente figura vemos un ejemplo extremo de un modelo en el que se observa un patrón extraño en los residuos que revela que algo no está funcionando en el modelo estimado

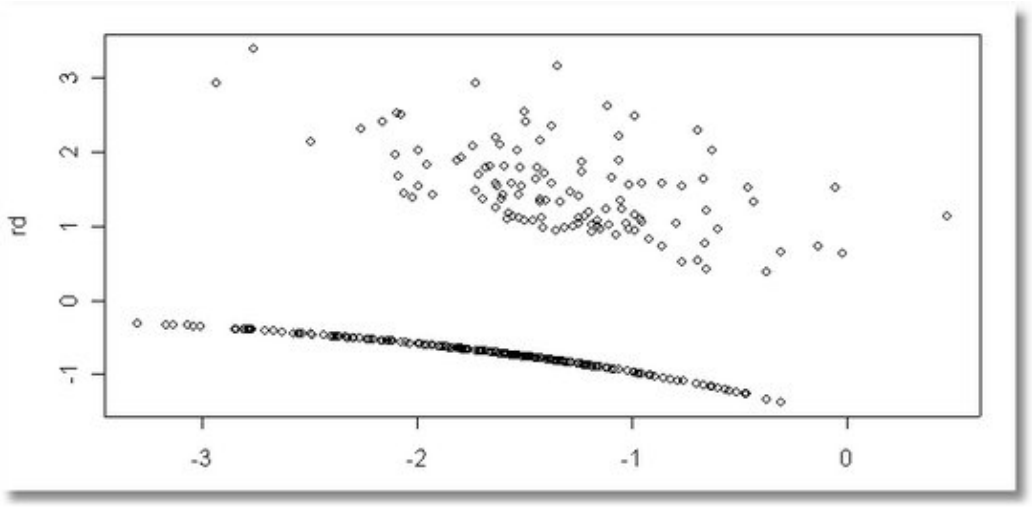

Fig. 2 Residuos de desviación frente a la puntuación de riesgo βX para cada paciente

Otro tipo de residuos que se emplean para verificar el modelo de regresión de Cox son los denominados **residuos de Schoenfeld**, siendo éstos los más efectivos en cuanto a detectar anomalías para cada una de las variables que intervienen en el modelo, sugiriéndonos por ejemplo que es utilizar alguna transformación para los datos. En el caso de los residuos de Schoenfeld tenemos un residuo para cada variable y para cada paciente, es decir que si tenemos un modelo de Cox con tres factores pronóstico se calcularán 3 residuos de Schoenfeld por paciente. Estos residuos valen cero para las observaciones incompletas, por lo que para facilitar su interpretación se suelen presentar en las salidas de ordenador sólo para los pacientes fallecidos. Es posible modificar estos residuos con el fin de que no valgan cero para las observaciones incompletas, obteniéndose entonces los denominados **residuos Schoenfeld corregidos o escalados**.

Verificación de la hipótesis de riesgos proporcionales

Una de las principales hipótesis del modelo de Cox es precisamente que la función de riesgo es proporcional dados dos perfiles de factores pronóstico distintos, y por tanto se debe mantener a lo largo del tiempo. Esto es algo que podemos verificar también en las gráficas de residuos.

Para facilitar la interpretación de estos gráficos se suele superponer una curva de ajuste, utilizando alguna función de ajuste local, de "*alisado*", que suelen estar disponibles en la mayor parte de los programas estadísticos, del tipo *ajuste por splines* o también como gráficas tipo *LOWESS* o *LOESS*.

En la figura 3 vemos un ejemplo de esta representación gráfica obtenida para los mismos datos y modelo que se representaba en la figura 2. En este caso se trata de los residuos Schoenfeld para uno de los factores pronóstico del modelo que es la EDAD en función del tiempo de supervivencia, y se ha ajustado una curva por el sistema de alisado por splines, junto con dos líneas adicionales a ±2 error estándar. Si se cumple la hipótesis de riesgos proporcionales los residuos debieran agruparse de forma aleatoria a ambos lados del valor 0 del eje Y, y la curva ajustada debería ser próxima a una línea recta.

Aquí vemos que el efecto de la edad disminuye con el tiempo, lo que está en contradicción con la hipótesis de riesgo constante a lo largo del tiempo de un modelo de Cox correcto.

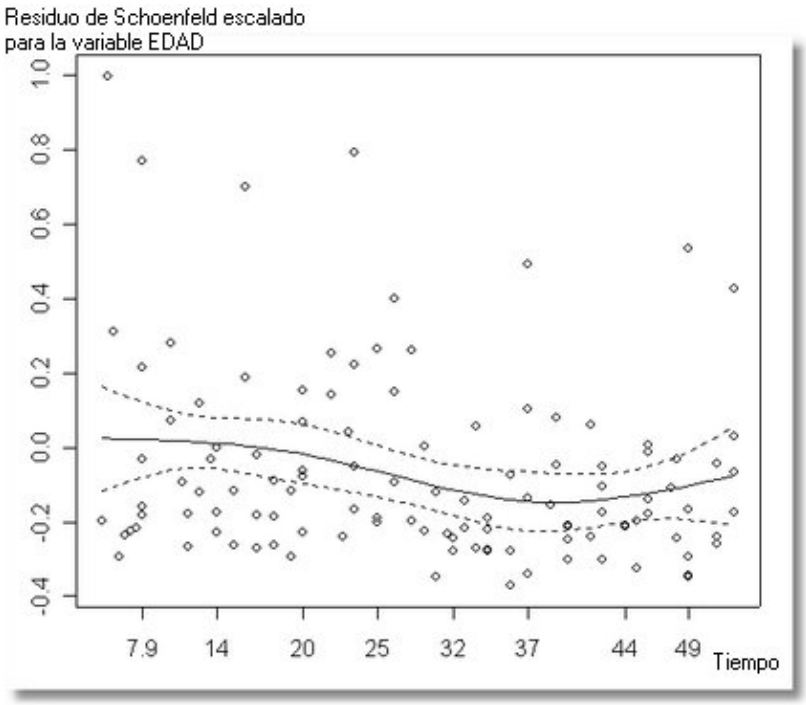

Fig. 3 Gráfico de los residuos de Schoenfeld escalados de una de las variables que intervienen en el modelo, en función del tiempo

Otra forma de verificar de forma gráfica la hipótesis de riesgos proporcionales en el modelo de Cox cuando la variable es cualitativa, consiste en representar ln(−lnS(t)) en función del ln(Tiempo) para cada uno de las categorías. Si se cumple la hipótesis de riesgos proporcionales éstas curvas tienen que ser aproximadamente paralelas. Si se analiza una variable numérica habrá que estratificarla previamente.

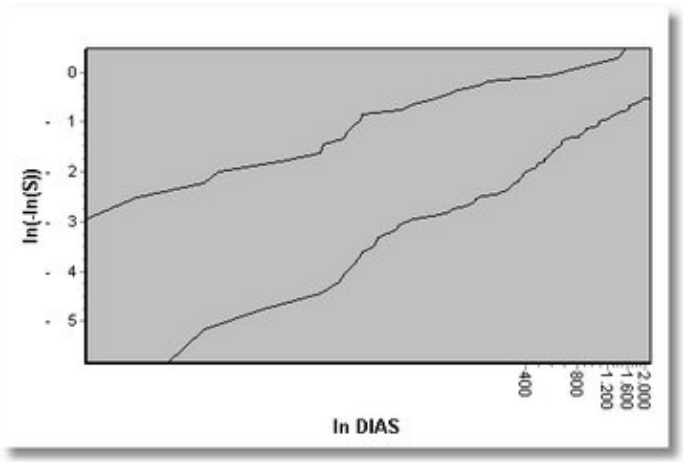

Fig. 4 Gráfica para verificar la hipótesis de riesgos proporcionales en dos grupos de pacientes

Lo dicho anteriormente se deduce debido a que la función de riesgo acumulada viene dada por la expresión

$$H_i(t) = H_0(t)\exp(\mathbf{b}'\mathbf{x}_i) \qquad [3]$$

por lo tanto tomando logaritmos y teniendo en cuenta la relación entre la supervivencia y la función de riesgo acumulada tenemos

$$H(t) = -\ln S(t) \qquad [4]$$

$$\ln H_i(t) = \ln H_0(t) + \mathbf{b}'\mathbf{x}_i \qquad [5]$$

$$\ln(-\ln S_i(t)) = \ln(-\ln S_0(t)) + \mathbf{b}'\mathbf{x}_i \qquad [6]$$

luego según la fórmula [6] las curvas en cada grupo seguirán la forma de la supervivencia base S_0 y se mantendrán paralelas de forma aproximada, separadas por la distancia marcada por el coeficiente b.

Si no se puede asumir que es correcta la hipótesis de riesgos proporcionales, una alternativa consiste en incluir en el modelo un elemento de interacción entre esa variable y el tiempo. Si para simplificar sólo tenemos la variable x1, la función de riesgo quedaría formulada de la siguiente manera:

$$h_i(t) = h_0(t)\exp(\beta_1 x_{1i} + \beta_2 x_{1i} t) \qquad [7]$$

donde interviene el producto X1.T que es por tanto un covariante dependiente del tiempo.

Otra alternativa cuando no se cumple la hipótesis de riesgos proporcionales es construir un modelo en el que la función de riesgo base $h_0(t)$ pueda variar de forma diferente en cada grupo (en el caso de que la variable sea numérica habría que estratificarla).

Y una tercera alternativa sería utilizar un modelo distinto de los modelos de Cox, pero ahí entramos ya en otro tema...

Observaciones que más influyen en la estimación de los coeficientes

Otra salida que suelen proporcionar los programas modernos de análisis de estadística es un vector de parámetros denominados **delta−beta**, que permiten investigar si alguna observación (algún paciente en particular) tiene un impacto importante en la estimación de los coeficientes de regresión. Una posible forma de averiguarlo consistiría, si tenemos N pacientes, en estimar el modelo N veces eliminando en cada ocasión uno de los pacientes, y evaluar el efecto que tiene en los parámetros estimados dicha eliminación. Si calculamos para cada variable incluida en el modelo y para cada paciente la diferencia entre el valor del coeficiente para esa variable en el modelo cuando intervienen todos los pacientes y el valor cuando se excluye el paciente, obtenemos los valores de delta−beta. Así para el paciente i el valor deltabeta correspondiente a la variable *j* es:

$$\mathrm{delta}_i\mathrm{beta}_j = \beta_j - \beta_j(\text{excluyendo } i) \qquad [8]$$

Conviene representar los valores de delta−beta para cada variable del modelo frente a los índices de paciente (número de orden), así como frente al tiempo de supervivencia, lo que nos permitirá detectar la posible presencia de observaciones con una influencia excesiva en el modelo, observaciones que convendrá investigar.

Los valores delta−beta pueden estandarizarse dividiéndolos por el error estándar del coeficiente correspondiente.

Epílogo

La elaboración de modelos de regresión para datos de supervivencia es un proceso complejo debido fundamentalmente a la presencia de observaciones incompletas o censuradas (pacientes para los que no se conoce su tiempo de vida) y por ello la verificación del modelo es bastante más complicada que en el caso de la regresión lineal. Previamente se ha debido proceder a una cuidadosa elección de las variables, así como a una buena revisión de los datos.

Conviene también tener presente que las ideas deben ser previas a la estadística, y los modelos matemáticos no suelen ser un buen sistema de generar hipótesis sino tan solo una herramienta para verificarlas.

Aunque la hipótesis principal en la que se basan los modelos de Cox es que el riesgo se mantiene proporcional en el tiempo y por lo tanto es una de las primeras cosas verificar, no debemos olvidarnos de otros aspectos, que aunque sean de sentido común conviene recordar. Es obvio que las causas deben preceder a los efectos y que por tanto no podemos utilizar datos futuros para predecir sucesos futuros. Así por ejemplo, supongamos que se dispone de un dato analítico como la creatinina, y se comprueba que es un factor pronóstico adecuado de la supervivencia; sería absurdo incluir en el modelo el valor medio de la creatinina durante el tiempo de seguimiento, o la diferencia entre el valor al comienzo del estudio y al cabo de un tiempo determinado. Sólo podemos utilizar los valores al comienzo del estudio.

Enlaces

- Regresión de Cox: Selección de Variables y Análisis de Residuos
  Técnicas estadísticas en el análisis de la supervivencia.
  Curso impartido en el Hospital de Sagunto sobre las técnicas básicas de supervivencia (Kaplan−Meier y comparación de curvas) y la regresión de Cox.
  José D. Bermúdez
  Depto. Estadística e I.O. Universitat de València. Noviembre de 1997
- Tests of proportional hazards in Cox modelling in SAS, STATA, R and SPLUS
  Preparado por UCLA Academic Thecnology Services

Bibliografía

- **Modelling Survival Data in Medical Research**
  David Collet
  Chapman&Hall/CRC. 2003

Indice de artículos

Principio de la página