VERIFICACIÓN DE LOS SUPUESTOS DEL MODELO DE COX

Rafael E. Borges P.
Escuela de Estadística,
Universidad de Los Andes,
Mérida 5101, Venezuela.
e-mail: borgesr@ula.ve

Temática: Métodos Estadísticos en Epidemiología.

Resumen

El modelo de Cox es el modelo de regresión para datos de sobrevida más utilizado en el área médica. Una práctica común ha sido la utilización del modelo de Cox sin la correspondiente verificación de sus supuestos. El desarrollo del enfoque basado en procesos de conteo ha permitido ampliar el espectro del análisis de supervivencia, y muchas de las técnicas han sido incorporadas a los paquetes estadísticos comerciales en los últimos años. En el presente trabajo, se presenta la metodología que debe seguirse para la verificación del supuesto de riesgo proporcional y el análisis de residuos, utilizado para estudiar la influencia de individuos en la estimación del modelo, así como de sus coeficientes y para la verificación de la adecuacidad de la forma funcional de las covariables continuas. Se presentan dos ejemplos: uno con datos de diálisis renal y otro con datos de cáncer de mamas.

1. Introducción.

En 1972, Cox introduce el modelo de regresión más utilizado en análisis de supervivencia, este modelo puede escribirse mediante:

$$\lambda\left(t;Z_i\left(t\right)\right)=\lambda_0\left(t\right)e^{\beta' Z_i(t)}$$

donde $Z_i\left(t\right)$ es el vector de covariables para el i-ésimo individuo en el tiempo *t*.

Este modelo incluye una parte paramétrica $r_i\left(t\right)=e^{\beta' Z_i(t)}$, llamada puntaje de riesgo y otra parte no paramétrica $\lambda_0\left(t\right)$, llamada función de riesgo base..

El modelo de regresión de Cox se llama también modelo de riesgos proporcionales debido a que el cociente entre el riesgo para dos sujetos con el mismo vector de covariables es constante en el tiempo, es decir:

$$\frac{\lambda\left(t;Z_i\left(t\right)\right)}{\lambda\left(t;Z_j\left(t\right)\right)}=\frac{\lambda_0\left(t\right)e^{\beta' Z_i(t)}}{\lambda_0\left(t\right)e^{\beta' Z_j(t)}}=\frac{e^{\beta' Z_i(t)}}{e^{\beta' Z_j(t)}}$$

A pesar de su amplia difusión, la verificación de los supuestos de los modelos y el análisis de los residuos, sólo ha estado disponible desde hace unos pocos años como producto del desarrollo del enfoque de análisis de supervivencia basado en procesos de conteo y

martingalas (Andersen et al., 1993, Fleming y Harrington, 1991, Therneau y Grambsch, 2000, Therneau et al., 1990).

Existen cuatro tipos de residuos de interés en el modelo de Cox: (i) los residuos de martingala, utilizados para verificar la forma funcional de un predictor continuo, (ii) los de desvíos (deviances), usados para la detección de valores atípicos (outliers), (iii) los de puntaje (score), utilizados para verificar la influencia individual y para la estimación robusta de la varianza y (iv) los de Schoenfeld, que se usan para la verificación del supuesto de riesgos proporcionales. En esta ponencia se presentan los residuos que se utilizan para la verificación de los supuestos del modelo de Cox y se presenta la verificación para un par de situaciones prácticas.

2. Metodología.

2.1. Residuos para la verificación de los supuestos.

2.1.1. Residuos de martingala.

Los residuos de martingala se definen como:

$$\hat{M}_i(t) = N_i(t) - \hat{E}_i(t) = N_i(t) - \int_0^t Y_i(s) e^{\beta' Z_i(s)} d\hat{\Lambda}_0(\beta, s)$$

donde $\hat{\Lambda}_0(\beta, s)$ es el estimador del riesgo base de Breslow (o de Tsiatis o de Nelson y Aalen) definido como:

$$\hat{\Lambda}_0(\beta, s) = \int_0^t \frac{\sum_{i=1}^n dN_i(s)}{\sum_{i=1}^n Y_i(s) e^{\beta' Z_i(s)}}$$

y están basados en la martingala de un proceso de conteo para el i-ésimo individuo, $M_i(t) = N_i(t) - E_i(t)$, definida mediante:

$$M_i(t) = N_i(t) - \int_0^t Y_i(s) e^{\beta' Z_i(s)} \lambda_0(s) ds$$

Los residuos de martingala son muy asimétricos y con una cola muy larga hacia la derecha, particularmente para datos de supervivencia para un solo evento.

2.1.2. Residuos de desvíos (deviances).

Los residuos de desvíos se obtienen mediante una transformación de normalización de los desvíos de martingala y son similares en forma a los residuos de desvíos (deviances) en la regresión de Poisson.

Los residuos de desvíos se definen de la manera siguiente: si todas las covariables son fijas en el tiempo, los residuos toman la forma:

$$d_i = \text{signo}\left(\hat{M}_i\right) * \sqrt{-\hat{M}_i - N_i \log\left(\left(N_i - \hat{M}_i\right)/N_i\right)}$$

Una expansión de Taylor de un término muestra que:

$$d_i \approx \frac{N_i - \hat{E}_i}{\sqrt{\hat{E}_i}}$$

que es formalmente equivalente a los residuos de Pearson de los modelos lineales generalizados.

2.1.3. Residuos de puntajes (scores).

Los residuos de puntajes se definen como:

$$U_{ij} = U_{ij}\left(\hat{\beta}, \infty\right)$$

donde $U_{ij}\left(\beta, t\right), j = 1, \cdots, p$ son las componentes del vector fila de longitud *p* obtenido a través del proceso de puntaje para el i-ésimo individuo:

$$U_i\left(\beta\right) = \int_0^t \left[Z_i\left(t\right) - \bar{Z}\left(\beta, t\right)\right] dN_i\left(t\right)$$

2.1.4. Residuos de Schoenfeld.

Los residuos de Schoenfeld se definen como la matriz:

$$s_{ij}\left(\beta\right) = Z_{ij}\left(t_i\right) - \bar{Z}_j\left(\beta, t_i\right)$$

con una fila por muerte y una columna por covariable, donde *i* y t_i son los individuos y el tiempo de ocurrencia del evento respectivamente.

2.2. Análisis de supervivencia y verificación de supuestos.

2.2.1. Caso 1: Análisis de supervivencia para datos de diálisis peritoneal.

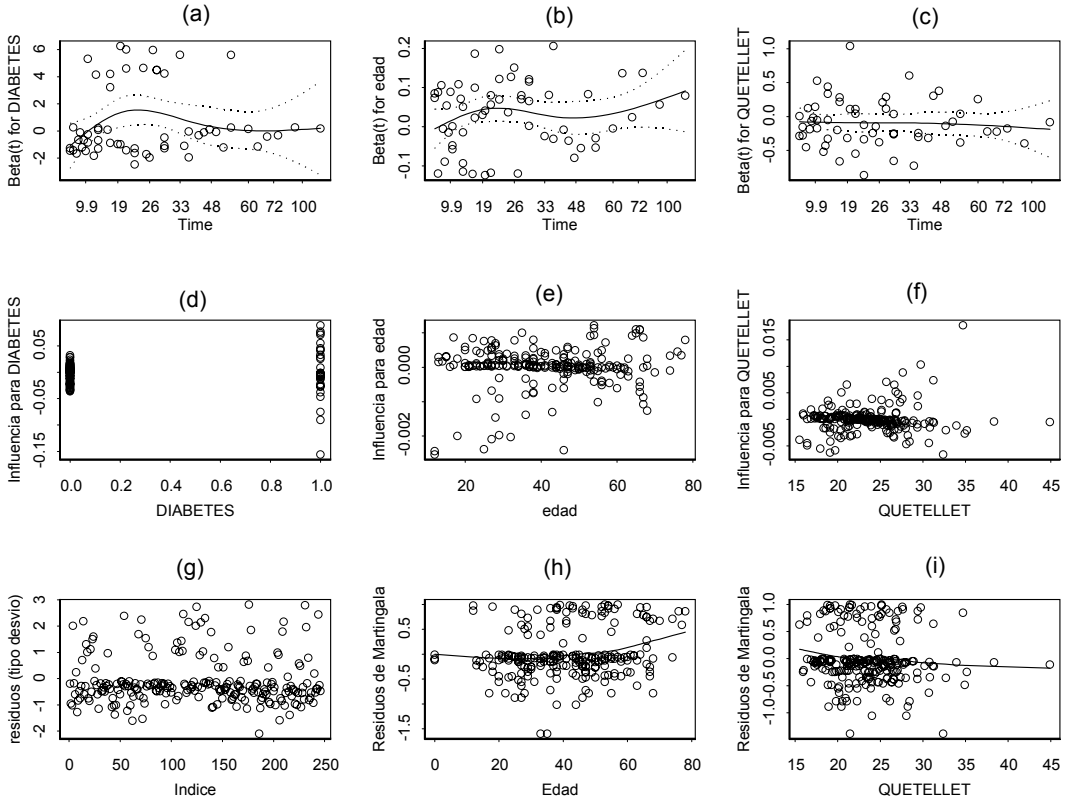
En 2002, Borges presenta los resultados de un análisis de supervivencia llevado a cabo con 246 pacientes en diálisis peritoneal (DPA) que acudían al Servicio del Hospital Clínico Universitario de Caracas entre 1980 y 1997. Los pacientes fueron seguidos desde el comienzo de sus sesiones de diálisis hasta alcanzar el evento de interés: Muerte por causas asociadas a la diálisis. Debido a que en no todos los pacientes se les observó alguno de estos dos eventos, algunas observaciones son censuradas. En el análisis se incluyeron 100 covariables dicotómicas y 16 continuas.

Al efectuar el ajuste de los modelos de Cox, se encontró que el mejor modelo era el que incluía como covariables significativas a: Diabetes, Edad e índice de Quetelet, todas significativas al 10%. Adicionalmente, el modelo resultó significativo al 0.1% por los tres criterios utilizados en el trabajo (razón de verosimilitud, Wald y puntajes). Las estimaciones de los coeficientes y sus interpretaciones pueden verse en Borges, 2002.

Sin embargo, todavía falta verificar los supuestos del modelo y esto se hará mediante técnicas gráficas utilizando el S-PLUS versión 6.1 (Insightful Corporation., 2001) y mediante la metodología desarrollada por Therneau (Therneau y Grambsch, 2000).

El gráfico No. 1 presenta las salidas que se utilizan para efectuar el análisis de residuos.

Gráfico No.1. Verificación de los supuestos del modelo de Cox.



Un análisis de los componentes del gráfico anterior nos dice lo siguiente:

Verificación de los supuestos de riesgos proporcionales: La verificación de los supuestos de riesgos proporcionales puede verse mediante (a), (b) y (c). En estos gráficos no se observa una violación del supuesto en cada una de las covariables.

Influencia de individuos en la estimación de los coeficientes: La contribución de los individuos en la estimación de los coeficientes puede verse a través de los gráficos (d), (e) y (f). En estos gráficos se puede observar que para diabetes edad no existen individuos que estén influyendo en la estimación de sus respectivos coeficientes. Para el índice de Quetelet, se observa que existe un individuo que probablemente esté influyendo en la estimación de su coeficiente, el que está ubicado en la parte superior, que corresponde al individuo # 6 de la base de datos.

Influencia de individuos en la estimación del modelo: En el gráfico (g) no se observa ningún individuo que esté influyendo en la estimación del modelo.

Forma funcional de las covariables continuas: En los gráficos (h) e (i) se observa que la forma funcional es correcta en el modelo, tanto para la edad, como para el índice de Quetelet. Eso quiere decir, que no hace falta que se efectúen transformaciones de estas covariables.

2.2.2. Caso 2: Análisis de supervivencia para datos de cáncer de mamas.

Actualmente se está analizando una base de datos que contiene el seguimiento de un grupo de 56 pacientes con cáncer de mamás servicio de Ginecologia y Anatomía Patológica del Instituto Autónomo Hospital Universitario de los Andes (I.A.H.U.L.A.), Mérida, Venezuela que ingresaron al servicio entre 1990 y 1994. Se espera presentar resultados preliminares del estudio en el simposio.

3. Conclusiones.

Al igual que como sucede en cualquier situación en donde se ajuste un modelo estadístico a un conjunto de datos, la verificación de los supuestos es de vital importancia para que el modelo pueda ser interpretado con validez. En el caso particular del modelo de Cox la verificación de los supuestos está disponible a través del software comercial y debería efectuarse a la hora de estudiar la calidad del ajuste del modelo.

4. Bibliografía.

Andersen, P.K., Borgan, ∅., Gill, R.D. y Keiding, N. (1993). *Statistical Models Based on Counting Processes*. N.Y.: Springer-Verlag.

Borges, R.E. (2002). *Análisis de Supervivencia Aplicado a un Caso de Diálisis Renal: Diálisis Peritoneal en el Hospital Clínico Universitario de Caracas y Hemodiálisis en el Hospital de Clínicas Caracas, 1980-2000*. Tesis de M.Sc. en Estadísitica Aplicada, Mérida, Venezuela: Instituto de Estadística Aplicada y Computación, Universidad de Los Andes. (Disponible en: http://tesis.saber.ula.ve)

Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*: Series B*,* **34:** 187-220.

Fleming, T.R. y Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. N.Y.: John Wiley & Sons, Inc.

Insightful Corporation (2001). *S-PLUS 6 for Windows Guide to Statistics, Volume 2*. Insightful Corporation, Seattle, WA.

Therneau, T.M. y Grambsch, P.M. (2000). *Modeling Survival Data: Extending the Cox Model*. N.Y.: Springer-Verlag.

Therneau, T.M., Grambsch, P.M. y Fleming, T.R. (1990). Martingale-based residuals for survival models. *Biometrika,* **77:** 147-160.