Análisis de supervivencia

Tarea 2

Fecha de entrega: 3 de octubre

1. Considera la base de datos *Turnover* para la cual se tiene el tiempo que pasan los empleados en sus trabajos (*stag*) antes de renunciar o ser despedidos (*event*).

   (a) Escoge al menos una de las variables binarias:

      i. *gender*: sexo del empleado.

      ii. *head_gender*: sexo del supervisor del empleado.

      iii. *greywage*: salario completamente declarado ante las autoridades (white) o salario dividido entre un porcentaje declarado y otro no (grey).

   Grafica las curvas de Kaplan-Meier y realiza la prueba de log-rangos y de Peto-Peto. ¿Hay diferencia en la supervivencia para la(s) variable(s) que elegiste?

   (b) Escoge al menos una de las variables categóricas:

      i. *industry*: tipo de industria en la que trabaja.

      ii. *profession*: la profesión del empleado.

      iii. *traffic*: medio por el cual el empleado se enteró del trabajo i.e.: advertising (advert), referencia de un amigo que no es empleado del lugar (recNErab), referencia de alguien que trabaja en el lugar (referal), aplicación mediante una página web de trabajos (youjs), agencia de reclutamiento (KA), amigo del empleador (friends), el empleador lo contactó por un contacto (rabrecNErab), o el empleador lo contactó al ver el CV en un página de trabajos (empjs).

      iv. *way*: medio de transporte que utiliza el empleado para llegar al trabajo.

   Grafica las curvas de Kaplan-Meier y realiza la prueba de log-rangos y de Peto-Peto. ¿Hay diferencia en la supervivencia?

2. Considera la base de datos *Fumadores* la cual describe los tiempos (*ttr*) de recaída (*relapse*) para un grupo de personas aleatoriamente sujetas a dos tratamientos diferentes (*grp*: terapia triple (combination) o únicamente parche (patch)) y contesta lo siguiente a partir de las curvas de Kapla-Meier y la prueba de log-rangos y Peto-Peto.

(a) ¿Hay diferencia en la supervivencia por tratamiento?

(b) ¿Hay diferencia en la supervivencia dependiendo el tipo de fumador (*levelSmoking*)?

(c) ¿Hay diferencia en la supervivencia dependiendo de la raza (*race*)?

(d) ¿Hay diferencia en la supervivencia dependiendo del tipo de trabajo (*employment*: tiempo completo (ft), tiempo parcial (pt) u otro esquema (other))?

3. Considera la base de datos *Transplante médula ósea* la cual contiene las siguientes variables

   i. *grupo*: 1 - leucemia linfocítica aguda (ALL), 2 - riesgo bajo de leucemia mieloide aguda (AML) y 3 - reisgo alto de AML.

   ii. *tiempo*: tiempo a la muerte.

   iii. *status*: 1 - muerto, 0 - con vida.

   iv. *edadP*: edad del paciente.

   v. *edadD*: edad del donante.

   vi. *tiempoE*: tiempo de espera para recibir el transplante.

   vii. *FAB* (sistema de clasificación francés - americano - británico): 1 - Grado 4 o 5 AML, 0 - en otro caso.

   viii. *MTX* (metotrexato): 1 - sí, 0 - no.

   Ajusta un modelo de Cox con todas las covariables y analiza la calidad del ajuste a partir de los residuales que creas convenientes.

4. Considera la base de datos *Cáncer de laringe*, la cual contiene las siguientes variables

   i. *estado*: estado del cáncer (del 1 al 4).

   ii. *tiempo*: tiempo a la muerte en meses.

   iii. *edad*: edad al momento del diagnóstico.

   iv. *anio*: año en el que se hizo el diagnóstico.

   v. *status*: 1 - muerte, 0 en otro caso.

   Ajusta un modelo de Cox con todas las covariables. ¿El supuesto de relación lineal se cumple para las variables continuas? Si no es el caso, ¿puedes sugerir una forma funcional que arregle el problema?. A partir de los residuales de Cox-Snell ¿se puede pensar que el modelo es adecuado?. Finalmente, ¿qué puedes decir del supuesto de riesgos proporcionales?

5. Considera la base *transplante de riñon*, la cual contiene las siguientes variables

   i. *id*: número de observación.

   ii. *tiempo*: tiempo a la muerte.

   iii. *status*: 1 - muerte, 0 en otro caso.

   iv. *sexo*: 1 - hombre, 2 - mujer.

   v. *raza*: 1 - blanco, 2 - negro.

   vi. *edad*: edad del paciente en años.

   Ajusta un modelo de Cox utilizando una interacción de sexo y raza tomando como grupo base el ser mujer blanca y realiza un análisis de devianza ¿consideras que hay valores influyentes en la base de datos? De ser así, retíralos y vuelve a ajustar el modelo ¿cómo se compara el ajuste completo contra el ajuste sin valores influyentes?

6. Suponga que se desea ajustar un modelo de riesgos proporcionales, esto es, el riesgo de morir al tiempo t del i-ésimo individuo es

$$h_i(t) \quad = \quad h_0(t)\exp(x^T\beta),$$

   donde se asume que el riesgo para el grupo base está dada por

$$h_0(t) \quad = \quad \lambda\gamma t^{\gamma-1}.$$

   (a) Deriva la función de supervivencia del i-ésimo individuo ¿cómo se distribuye?.

   (b) Asume que sólo se tiene una covariable que toma los valores de 0 o 1 (para comparar dos grupos). Bajo el supuesto de riesgos proporcionales ¿qué debe suceder si se grafica $\log(-\log S_i(t))$ como función de $\log t$ para $i = 0, 1$?

   (c) Deriva la log-verosimilitud bajo el supuesto de que en n individuos hubo m muertes y $n - m$ censuras.

   (d) Utilizando la función *optim* de R escribe una rutina numérica que te permita calcular los estimadores de $\beta_1, \ldots, \beta_p$, λ y γ para la base de *NCCTG Lung Cancer* utilizando las covariables de sexo y ECOG. Compara tus resultados con los que se obtienen de utilizar la función *flexsurvreg* de la paquetería *flexsurv*.

   (e) Obtén los residuales de Cox-Snell y los de devianza y realiza un análisis de residuales para verificar el ajuste del modelo.

Actividades de DataCamp

1. *Survival Analysis in R*

2. *Machine Learning for Marketing Analytics in R*