Modelos de supervivencia y de series de tiempo

Tarea 1

Fecha de entrega: 6 de marzo

Instrucciones. La tarea se puede realizar en equipos de máximo 3 integrantes. Se entregará física (o digital) a la hora de la clase y no habrá prórroga. Las actividades de DataCamp son individuales y se podrán terminar hasta el final de la fecha de entrega y contarán como un punto extra sobre la tarea. En caso de no realizarlos habrá una penalización de un punto sobre la tarea.

1. (1 punto) Sea T una v.a. no negativa y absolutamente continua con función de supervivencia $S(t)$. Demuestra que
$$\mathbb{E}(T) \quad = \quad \int_0^\infty S(t)dt.$$

2. (1 punto) Sea T una v.a. con distribución Weibull de parámetros λ, γ. Encuentra $\mathbb{E}(T)$.

3. (1 punto) Para la distribución log-logística considera la parametrización dada por
$$f(t \mid \beta, \alpha) \quad = \quad \frac{\left(\frac{\beta}{\alpha}\right)\left(\frac{t}{\alpha}\right)^{\beta-1}}{\left(1+(\frac{t}{\alpha})^{\beta}\right)^2}.$$
Encuentra: $F(t)$, $S(t)$, $h(t)$ y la mediana de la distribución.

4. (1 punto) Considera el modelo log-normal de parámetros μ, σ^2 y realice lo siguiente.
   (a) Encuentra la media y la mediana de la distribución.
   (b) Demuestra que para la función de riesgo $h(t)$ se cumple que
   $$\lim_{t\to 0} h(t) = 0 = \lim_{t\to\infty} h(t).$$
   ¿Qué puedes concluir de esta función de riesgo?
   (c) Para un valor de μ fijo (e.g. $\mu = 0$) grafica la función de riesgo para $\sigma^2 < 1$, $\sigma^2 = 1$ y $\sigma^2 > 1$. ¿Cómo se comporta la función de riesgo para tu elección de parámetros?

(d) Para un valor de σ^2 fijo (e.g. $\sigma^2 = 1$) grafica la función de riesgo para $\mu < 1$, $\mu = 1$ y $\mu > 1$. ¿Cómo se comporta la función de riesgo para tu elección de parámetros?

5. (1 punto) Considera el modelo gamma generalizado y realiza un programa en el software de tu elección que permita calcular su función de riesgo $h(t)$ y analiza el comportamiento de la función, para una λ fija, en los siguientes escenarios.

   (a) θ y ρ tales que $\theta > 1$ y $\rho\theta > 1$. ¿Qué comportamiento tiene $h(t)$?

   (b) θ y ρ tales que $\theta < 1$ y $\rho\theta < 1$. ¿Qué comportamiento tiene $h(t)$?

   (c) θ y ρ tales que $\theta < 1$ y $\rho\theta > 1$. ¿Qué comportamiento tiene $h(t)$?

   (d) θ y ρ tales que $\theta > 1$ y $\rho\theta < 1$. ¿Qué comportamiento tiene $h(t)$?

6. (1 punto) En este ejercicio se analizará la función de distribución de Gompertz y Makeham para la cual se tiene la densidad

$$f(t) = \lambda\phi^t \exp\left\{\frac{-\lambda}{\log(\phi)}(\phi^t - 1)\right\}.$$

   Encuentre la función de riesgo $h(t)$ y realice lo siguiente

   (a) Demuestre que para $\phi > 1$ se tiene que

$$\lim_{t\to\infty} h(t) = \infty.$$

   (b) Demuestre que para $\phi < 1$ se tiene que

$$\lim_{t\to\infty} h(t) = 0.$$

   Notando que $h(0) = \lambda$, ¿qué interpretación le puedes dar a la función de riesgo para los dos casos previamente analizados? ¿qué caso se recupera cuando $\phi \to 1$?

7. (2 puntos) Para la base de datos *NCCTG Lung Cancer* ajusta una distribución log-normal, gamma generalizada o Gompertz y compara tus resultados con el modelo exponencial y el modelo Weibull vistos en clase. ¿Cuál considerarías que es el mejor modelo para estos datos?

8. (1 punto) Utiliza la expansión de Taylor de $\exp(-x)$ para mostrar que el estimador Kaplan-Meier es una aproximación al estimador Nelson-Aalen. ¿Qué condiciones se le debe pedir a d_j con respecto a n_j? Más aún demuestra que el estimador de Nelson-Aalen siempre será mayor al de Kaplan-Meier para todo t.

9. (1 punto) Considera la base de datos *turnover.csv* para la cual se tiene el tiempo que pasan los empleados en sus trabajos (*stag*) antes de renunciar o ser despedidos (*event*).

(a) Encuentra el estimador de Kaplan-Meier para estos datos.

(b) A partir de este estimador, elige y ajusta el modelo paramétrico que consideres adecuado.

(c) Para tu modelo paramétrico estima y grafica $f(t)$, $h(t)$ y $S(t)$ y proporciona la mediana del tiempo antes de renuncia/despido junto con su respectivo intervalo de confianza al 99%.

Actividades de DataCamp

1. Working with Dates and Times in R
2. Survival Analysis in R