Notas

Adriana Pérez-Arciniega Soberón

2016

1 Notas Sesiones

Tenemos una muestra de microcostos de enfermedades crónicas de un cierto número de individuos los cuales tienen asociadas un número de covariables sociodemográficas, socioeconómicas y médicas. El objetivo es modelar la duración y el costo de estos padecimientos por individuo.

Sea y_i una variable aleatoria respuesta para cada individuo i. $i = 1, .., n$ Cada individuo i tiene asociada información adicional x_i, por lo que $x_i = x_{i1}, ..., x_{ip}$ pertenece a $\mathbb{R}^p$.
Necesitamos una función de distribución de la variable respuesta dado las covariables tal que el valor esperado sea una función h de las covariables individuales por β.
$y_i|x_i$ con función de distribución $F(y_i|x_i)$ tal que $\mathbb{E}(y_i|x_i) = h(x_i'\beta)$.

Tomemos y_i en $\mathbb{R}_+$ y $x_i'\beta$ es un proyector lineal: $\mathbb{R}^p \to \mathbb{R}$ por lo que se necesita $h(\cdot)$ tal que $h : \mathbb{R} \to \mathbb{R}_+$.
Por lo cual podemos definir $\mathbb{E}(y_i|x_i) = exp(x_i'\beta)$. Dado que como sabemos, la función exponencial siempre nos arrojará un resultado positivo. Y esto es un **modelo lineal generalizado**.

Si tenemos una serie de individuos observados por un período de tiempo con costos asociados a su padecimiento, tendríamos entonces dos variables $(d_i, c_i)_{i=1}^{n_T}$ con d_i como duración y c_i como el costo, estas asociados al individuo i. Así que este modelo estaría determinado por datos de duración y costos, que se modelaría como un proceso **Poisson Marcado** con:

- c_j's positivas.
- d_j's positivas.

Tomemos como primer acercamientos a las d_j's (duraciones) de un solo individuo. Podemos pensar en una función de densidad:
$f(d_i) = h(d_i)S(d_i)$
donde $h(d_i)$ es la función hazard y $S(d_i)$ es la función de supervivencia.
Ligando el proceso Markov marcado explorado anteriormente podríamos decir

que la duración actual depende de la duración anterior:
$f(d_i|d_{i-1}) = h(d_i|d_{i-1})S(d_i|d_{i-1})$.

Así, si tomamos un individuo i:
$f(d_{ji}|x_i) = h(d_{ji}|x_i)S(d_{ji}|x_i)$. Esto se define como un **Modelo de Cox**.
Entonces,
$f(d_{ji}|x_i) = h_b(d_{ji})\varphi(x_i'\beta)S(d_{ji})$
En el contexto de un **Proceso Markov Marcado**:
$f(d_{ji}|d_{j-1,i}, x_i) = h_b(d_{ji}|d_{j-1,i})\varphi(x_i'\beta)S(d_{ji})$
En donde:

- $h_b(\cdot)$ se le conoce como la función hazard base que es igual para cualquier duración de cualquier individuo, es conocido como **baseline model**.

- $\varphi(\cdot)$ es fija y conocida como la función exponencial y son comunes a **todas** las observaciones de cada individuo i.

- β es común para cualquier individuo i

Supongamos que n_j como el número de observaciones del individuo j y m el número de individuos en la muestra.

$$\begin{bmatrix} d_{11}c_{11} & d_{12}c_{12} & \dots & d_{1n}c_{1n} \\ d_{21}c_{21} & d_{22}c_{22} & \dots & d_{2n}c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1}c_{m1} & d_{m2}c_{m2} & \dots & d_{mn}c_{mn} \end{bmatrix}$$

La función que liga las duraciones del mismo individuo es la **función hazard** $h_b(\cdot)$, que podemos estimar empíricamente con el método Kaplan-Myer. Y la función que liga las duraciones entre individuos es la $\varphi(x_i'\beta)$. La β es el parámetro a estimar, pero es igual común para todos los individuos i.

Ya que tenemos las funciones comunes a todos los individuos, podemos hacer predicción con la entrada de cualquier individuo x_{i+1} en un tiempo fututo t. Entonces tendríamos

$$\mathbb{P}(d_{i+1,1}^t, d_{i+1,2}^t, ..., d_{i+1,n}^t) = h_b(d_{i+1}^t)\varphi(x_{i+1}^t\beta)\prod_{j=2}^{n_t} h(d_j^t|d_{j-1}^t)\varphi(x_{i+1}^t\beta)$$

.

2 Regression Models and Life-Tables, Cox y Autoregressive Conditional Duration, Engel y Russell

Si consideramos una población de individuos y para cada individuo observamos el "tiempo de falla" o el "tiempo de censura". Sea T una variable aleatoria que representa el tiempo de falla y definamos la función de supervivencia $\mathbb{F}(t)$:

$$\mathbb{F}(t) = \mathbb{P}(T \geq t)$$

y sea $\lambda(t)$ la función hazard o la tasa de de falla según la edad.

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{\mathbb{P}(t \leq T \leq t + \Delta t | t \leq T)}{\Delta t}$$

Es decir, que la función hazard es la función que nos dice la probabilidad de que el tiempo de falla suceda en el tiempo t.
Supongamos una población de n individuos y que cada uno de los individuos tiene p variables asociadas $z_1, ..., z_p$, así para el j-ésimo individuo tiene el vector $\mathbf{z}_j = (z_{1j}, ..., z_{pj})$, supongamos que las z's son asociadas al tiempo. Entonces, tendríamos una función hazard:

$$\lambda(t; z) = \exp(z\beta)\lambda_0(t)$$

En esta función β es un vector de parámetros desconocidos y $\lambda_0(t)$ es la función hazard para un conjunto de condiciones **z=0**. Ligándolo con lo anterior en **Notas de Sesiones**, podemos ver que $\lambda_0(t)$ es equivalente a la $h_b(\cdot)$ y la $exp(z\beta)$ es equivalente a $\varphi(x_i'\beta)$.
Tomemos los distintos tiempos de falla como $t_1 < ... < t_k$, sea m_i el número de fracasos en t_i tal que $\sum m_i = n$ y r_i el número de individuos en riesgo al tiempo t_i; entonces tendríamos un estimador de la función hazard:

$$\tilde{\lambda}(t) = \sum_{i=1}^{k} \frac{m_i}{r_i} \delta(t - t_i)$$

Esta fórmula sería análoga a lo que se presenta en Engel y Russell como el tiempo de ocurrencia modelando el tiempo entre eventos.

En el paper de Engel y Russell exploran un modelo donde el tiempo entre los eventos es tratado como un proceso estocástico y propone una nueva clase de proceso marcado con tasas de llegada dependientes. Como este modelo se enfoca en la duración esperada entre eventos, este modelo se llama **Autorregresión con Duración Condicional (ACD)**. Ell paper de Engel y Russell habla de microfinanzas y de cómo es imposible analizar el trading de alta frecuencia en un espacio de tiempo fijo, por lo que se propone tomar el tiempo de ocurrencia como variable aleatoria que sigue un proceso puntual; asociado al tiempo de ocurrencia hay otras variables aleatorias a las que llamamos **marcas**, en el caso

de microfinanzas estas marcas corresponderían al volumen, a la diferencia en el precio de oferta y demanda o el precio; en el contexto de la tesis, estas marcas sería el costo del padecimiento.
La intensidad condicional se parametriza en términos de eventos pasados, por lo que la formulación básica de este modelo es la dependencia de la intensidad condicional en duraciones pasadas, es por eso que lo llamamos el modelo de autorregresión con duración condicional. Asociado con la intensidad está la esperanza condicional del tiempo de espera al siguiente evento. Es importante notar que el modelo está formulado en **tiempo de ocurrencia** pero modela la frecuencia y distribución del tiempo calendario entre eventos; ¿en el paper de Cox este tiempo de ocurrencia sería el tiempo de falla?
Consideremos un proceso estocástico $\{t_0, ..., t_n\}$ con $t_0 < t_1 < ... < t_n$ que es una secuencia de tiempos de ocurrencia dependientes en el tiempo, asociado a estos tiempos de ocurrencia tenemos la función $N(t)$ que es el número de eventos que han ocurrido para el tiempo t, si hay características asociadas al tiempo de ocurrencia se le conoce como un **proceso puntual marcado**. Hay dos generalizaciones de un proceso puntual:

- Proceso puntual que evoluciona sin efectos secundarios:
  Si para cualquier $t > t_0$ la realización de puntos durante $[t, \infty)$ no depende de la secuencia de puntos en el intervalo $[t_0, t)$.

- Proceso puntual condicionalmente ordenado:
  Si en $t \geq t_0$ hay un intervalo de tiempo suficientemente corto y condicional a cualquier evento P definido por la realización del proceso en $[t_0, t)$, que la probabilidad de dos o más eventos es infinitesimalmente relativa a la probabilidad de un evento.

Para este trabajo nos concentraremos en procesos puntuales que evolucionen con efectos secundarios y que sean condicionalmente ordenados.
Una función de intensidad de un proceso puntual "self-exciting", esto es, un proceso donde el pasado impacta la posible estructura de eventos futuros tiene la siguiente forma:

$$\lambda(t|N(t), t_1, ..., t_{N(t)}) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(N(t + \Delta t) > N(t)|N(t), t_1, ..., t_{N(t)})}{\Delta t}$$

Hay dos maneras de interpretar esta función de intensidad:

- Por tiempo calendario:

$$\lambda(t|N(t), t_1, ..., t_{N(t)}) = \omega + \sum_{i=1}^{N(t)} \pi(t - t_i)$$

  Cada t_i que pasa contribuye con $\pi(t - t_i)$ a la intensidad del tiempo t, π es una medida infectológica. Este modelo tiene el efecto marginal de que un evento que ocurrió hace X tiempo es independiente de lo que sucede

sin importar cuántos eventos hayan sucedido.

- Por tiempo de ocurrencia modelando el tiempo entre eventos:

$$\lambda(t|N(t), t_1, ..., t_{N(t)}) = \omega + \sum_{i=1}^{N(t)} \pi_i (t_{N(t)+1-i} - t_{N(t)-i})$$

así el impacto de la duración entre eventos sucesivos depende del número de eventos que intervienen.

Este tipo de modelos es estudiado por Cox en el marco de modelo de hazard proporcionales, donde la función de intensidad condicional de este modelo sería la descrita al inicio de esta sección donde z_i es el vector de variables explicatorias asociadas al **tiempo de llegada** i y el tiempo de falla, en este caso, esta condicionado por el vector de variables asociadas al número de eventos ocurridos para el tiempo t.

$$\lambda(t|z_{N(t)}, ..., z_1) = \lambda(t) \exp(\beta' z_{N(t)})$$

El modelo ACD está especificado en términos de la densidad condicional de las duraciones. Sea $x_i = t_i - t_{i-1}$ el intervalo de tiempo entre dos realizaciones, es decir, la duración. La densidad de x_i condicionado en las x's pasadas sean especificadas directamente.
Sea ψ_i la esperanza de la duración i dado por:

$$E[x_i|x_{i-1}, ..., x_1] = \psi_i(x_{i-1}, ..., x_1) = \psi_i$$

Así, la esperanza condicional de la duración depende de las duraciones pasadas. Asumimos, también, que:

$$x_i = \psi_i \epsilon_i; \quad \epsilon_i \sim p(\epsilon, \phi)$$

p es cualquier función de densidad y ϕ tiene varianza libre.
Sea p_0 la función de densidad de ϵ y sea S_0 su función de supervivencia asociada. Definimos como riesgo basal:

$$\lambda_0(t) = \frac{p_0(t)}{S_0(t)}$$

$$\Rightarrow \lambda(t|N(t), t_1, ..., t_{N(t)}) = \lambda_0(\frac{t - t_{N(t)}}{\psi_{N(t)+1}}) \frac{1}{\psi_{N(t)+1}}$$

Este modelo es conocido como modelo de vida acelerada dado que la información pasada influencia el ritmo en el que el tiempo pasa. La velocidad del tiempo depende de los tiempos de ocurrencia pasados a través de la función ψ. La versión más simple del modelo ACD asume que las duraciones son exponenciales condicionalmente.

$$\lambda(t|x_{N(t)}, ..., x_1) = \psi^{-1}_{N(t)+1}$$

Con una memoria de m, la intensidad condicional implica que solo los momentos más recientes han influenciado la duración:

$$\psi_i = \omega + \sum_{j=0}^{m} \alpha_j x_{i-j}$$

En general,tomaremos el modelo ACD (m, q) que se refiere al orden de desfases, este modelo es conveniente porque permite calcular varios momentos.

$$\psi_i = \omega + \sum_{j=0}^{m} \alpha_j x_{i-j} + \sum_{j=0}^{q} \beta_j \psi_{i-j}$$

El modelo ACD es propuesto como un modelo con tiempos de ocurrencia correlacionados intertemporalmente. Para examinar la dependendia, calculamos las autocorrelaciones y las autocorrelaciones parciales en el tiempo de espera entre eventos.