

Survival analysis

(Besprechung: Mo., 14.01.2013)

1. Zur Erinnerung: *Linkszenisierte Daten lassen sich durch Standardverfahren für rechtszensierte Daten analysieren, wenn man die Zeitachse 'umdreht', indem man die transformierte Beobachtungszeiten $\tilde{t}_i = \tau - t_i$, $\tau > t_i$ für alle $i = 1, \dots, n$ definiert.*

Der Datensatz `baboon.txt`, den man auf der Homepage zur Vorlesung (unter 'Material') findet, enthält in einem Reservat in Kenia erhobene Daten über einen Pavianstamm. Die Variable `time` gibt an, zu welcher Tageszeit (gemessen in Minuten) die Paviane den Baum, in dem sie schlafen, verlassen. An einigen Tagen war dies bereits der Fall, bevor die Beobachter eingetroffen waren, so dass die Daten linkszensiert sind. Der Indikator `observed` gibt an, ob die exakte Zeit beobachtet wurde (`observed=1`) oder nicht (`observed=0`). Schätzen Sie die Verteilung des Zeitpunkts zu dem der Baum verlassen wird, indem Sie die Zeitachse um Mitternacht, $\tau = 1440$, umkehren und für die transformierten Daten den Kaplan-Meier-Schätzer und die zugehörige Verteilungsfunktion berechnen.

Berechnen Sie anschließend den Kaplan-Meier-Schätzer und die zugehörige Verteilungsfunktion für die Originaldaten und stellen Sie die Schätzer für die transformierten Daten und die originalen Daten in einer vergleichenden Grafik dar.

(4 Punkte)

2. In dieser Aufgabe benutzen wir Daten aus einer Studie über eine Intrauterinpeessare (Spirale) zur Empfängnisverhütung (Peterson, 1975). Hundert Frauen haben an dieser Studie teilgenommen, und sie wurden bis maximal 300 Tage beobachtet. Für jede Frau wurde die Zeit (in Tagen) ab der Implantation bis zum ersten der folgenden Ereignisse notiert:

- Die Spirale wird ausgestoßen.
- Die Spirale wird entfernt.
- Die Frau nimmt nicht mehr an der Studie teil (Zensierung).

Folgende Daten wurden aufgenommen:

Ausgestoßen	Entfernt		Zensierung
2	14	166	86
8	21	178	203
10	27	183	207
25	40	272	
28	42	272	
28	92	288	
32	110	288	
63	147	288	
86	148	297	
158	165		

- (a) Beschreibe Risiken und Zensierungen bei der Studie.

- (b) Bestimme, für alle die Ereigniszeiten t_i , die Werte von R_i , D_i und Y_i aus der Vorlesung.
- (c) Berechne die geschätzte kumulative Inzidenzfunktionen, $\widehat{CI}_1(t_k)$ und $\widehat{CI}_2(t_k)$ für $K = 1, \dots, 4$.
- (d) Definiere eine Funktion in R, die die kumulative Inzidenzfunktion berechnet (gegeben t_i, R_i, D_i und Y_i). (oder, wenn man will, gegeben, t_i, R_i, D_i und Y_i und $\widehat{S}_{KM}$, der KM-Schätzer für die gemeinsame Survivalfunktion). Berechne die geschätzte kumulative Inzidenzfunktion für die zwei Risiken und stelle sie graphisch dar (der Befehl `stepfun` könnte nützlich sein).
- (e) Vergleiche (graphisch) $\widehat{CI}_1(t)$:
- mit der Schätzung der Survivalfunktion, die man bekommt, wenn man nur die 'Ausstoßungen' als Ereignis betrachtet (und die 'Entfernungen' wegläßt);
 - und mit der Schätzung der Survivalfunktion, die man bekommt, wenn man die 'Entfernungen' als Zensierungen betrachtet (und die Ausstoßungen als Ereignis).

Was bedeuten die drei Kurven?

- (f) Es gibt auch eine fertige R-Funktion, um die Kumulative Inzidenzfunktionen zu rechnen: `cuminc`, aus dem Paket `cmprsk`. Die Funktion braucht als Input eine Variable, bei welcher die Ereignisse eine unterschiedliche Kodierung nach Ausfallsgrund haben (man könnte einen Vektor angeben, der den Wert 0 bei Zensierungen, 1 bei Ausstoßungen und 2 für Entfernungen hat). Vergleiche graphisch die Kurven, die man auf dieser Art bekommt mit denen, die man davor gerechnet hat.

(8 Punkte)