Rによる生存分析について

2004MM090 山口暁
指導教員: 田中豊

1　はじめに

生存分析は、あるイベント（故障、死亡など）が起きるまでの時間とイベントの間の関係に焦点を当てる分析方法であり、工学分野においては機械システムや製品の故障など、医学分野においては、疾患の病気の再発や死亡などを対象とした研究分野である。本研究では、医学分野にしぼり、また生存分析は共変量（年齢や性別などの複数の要因、説明変数）を導入するか、生存時間の分布形に特定の確率分布を仮定するかによって「ノンパラメトリックモデル」「セミパラメトリックモデル」「パラメトリックモデル」の3種類に分けられるので、それぞれのモデルについて理解し、実際にデータに適用してその機能を検討していく。

2　生存分析の基本概念

2.1　生存関数とハザード関数

生存時間Tを累積確率分布関数$F(t)$と確率密度関数$f(t)$に従う非負の確率変数とする。イベントがある地点tまで起きていない生存関数$S(t)$は

$$S(t) = 1 - F(t) \tag{1}$$

と表わされる。また、イベントがある地点tまでに起きていないという条件の下で、次の瞬間にイベントが起こる瞬間死亡率を示すハザード関数$h(t)$は

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{2}$$

と表わされる。そして、生存関数$S(t)$とハザード関数$h(t)$の関係式は

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \\ &= \frac{d}{d(t)} \times (-\log S(t)) \end{aligned} \tag{3}$$

と表わされる。従って、生存関数$S(t)$とハザード関数$h(t)$と累積ハザード関数$H(t)$の関係式は

$$H(t) = \int_0^t h(t)dt = (-\log S(t)) \tag{4}$$

と表わされる。

2.2　打ち切り

調査対象者が、調査中になんらかの理由でいなくなったり、調査目的とは別の目的で死亡してしまい、途中でデータが取れなくなることがある。これを「打ち切り」と言い、それを考慮して考える必要がある。

3　ノンパラメトリックモデル

3.1　カプラン・マイヤー推定法

ノンパラメトリックモデルは、確率分布関数を指定しないで生存分析を推定するモデルである。代表的な推定方法は、カプラン・マイヤー推定法である。死亡（イベント）があった時点をt_1,t_2とし、t_1時点での死亡数をd_1、t_2時点での死亡数をd_2として、以下同様であるとする。さらに$t_1, t_2, \ldots$ の直前のリスク集合の大きさ（その直前までまだ生存していた個体の数）を$n_1, n_2, \ldots$で示す。

$$\begin{aligned} \hat{S}(t) &= (1 - \frac{d_1}{n_1}) \times (1 - \frac{d_2}{n_2}) \times \ldots \\ &= \prod_{t_i \leq t} (1 - \frac{d_i}{n_i}) \end{aligned} \tag{5}$$

と表わされ、これは、死亡が起きていない時点のハザードを0とし、死亡が起きている時点のハザード成分を（死亡数/リスク集合）として、(1−ハザード成分)を時点tまでの時点について掛け合わせた関数である。 信頼区間の代表的な計算方法は、$\hat{S}exp[\pm z_{1-a/2} se(\hat{H})]$ である。なお、式の中の$\hat{S}$は生存関数の推定値$\hat{S}(t)$、$\hat{H}$はハザード関数の推定値$\hat{H}(t)$、seは標準誤差を示す。

3.2　ログランク検定

単一の対象者集団の生存時間を単に調べるのではなく、2つの異なる集団の生存時間の分布を比較したい時がある。この目的の一つは、生存状況が群間で系統的に異なるかどうかを調べるためである。もし、データに打ち切りがある場合には、普通の比較ができない。そういう時に用いられる方法の1つがログランク検定である。ログランク検定は、死亡というイベントが発生した地点tごとに、グループと生存状況を示す2×2分割表を構成し、その結果を$Mantel - Haenszel(1959)$の考え方で総合して群間の差を検定しようとするアイデアがもとになっている。

4　セミパラメトリックモデル

セミパラメトリックモデルは、共変量を考慮するモデルである。セミパラメトリックモデルの代表的なものとしてコックス比例ハザードモデルがある。これは、共変量$x = (x_1, x_2, \ldots, x_m)^T$ を持つハザード関数を$h(t|x)$とする。また、xを変数とする関数を$h(t|x)$とし、あるハザード関数を$h_0(t)$とすると、次の式で定義される。このモデルは、生存時間を目的変数とした回帰モデルである。

$$h(t|x) = h_0(t)exp(\beta_x) = h_0(t)exp(\sum_{i=1}^{m} \beta_i x_i) \tag{6}$$

5　パラメトリックモデル

パラメトリックモデルは、共変量を導入し、生存時間が確率分布に従うモデルのことである。また生存時間モデルに多く使われる確率分布は、指数分布とワイブル分布である。パラメトリックモデルの特徴としては、コックス比例ハザードモデルと比べ、計算速度は速いが、生存時間の確率分布を仮定する制約条件があるので、応用範囲が狭いと言われている。

6　データ解析

使用するデータは、統計ソフトRに組み込まれているポータブル透析装置の使用と腎臓患者の生存に関して、38ペア（使用と不使用）76人に対する臨床試験データ$kidney$を用いる。まず、カプラン・マイヤー推定法によるデータ解析を行った。

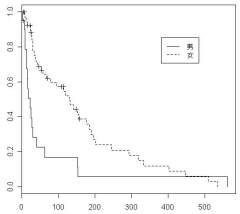

図 1: 性別と生存時間との比較

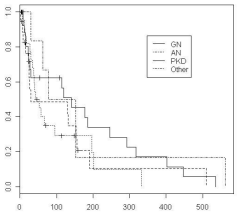

図 2: 病気の種類と生存時間との比較

カプラン・マイヤー推定法でのデータ解析結果として、性別と生存時間との比較の生存曲線を図1に、病気の種類と生存時間との比較の生存曲線を図2にそれぞれ示した。また、ログランク検定を行ってみると、性別に関して、$p = 0.0324$となり、有意であることが分かった。

次にコックス回帰によるデータ解析を行った。

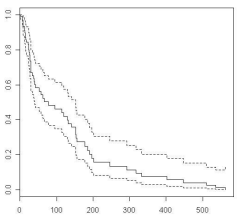

図 3: 生存曲線

コックス回帰では、病気の種類と性別が生存時間にどのように影響を与えているかを調べた。結果として、図3に生存曲線を示した。また$summary$表の結果から、性別のp値が$p = 0.000035$で、$disease$のp値が$P = 0.02$より、性別と$disease$が有意であった。また、$disease$の間の差について尤度比検定とワルド検定を行ったところ、尤度比検定の結果が、$p = 0.0147$で、ワルド検定の結果が、$p = 0.0315$より有意であることが分かった。

7　シミュレーション

7.1　人口データの作成方法

データ解析時に用いたデータでは、変数に偏りがあり、正確な結果が出せたとは言えない。そこで、乱数を使い、変数の偏りをなるべく無くして解析をする。作成方法としては、一般的に、生存曲線は指数分布に従うとされているので、平均の違う2群のデータを作成する。そこで、指数乱数をそれぞれ100個ずつ2群発生させ、どのデータを打ち切るか、またどこで打ち切るかを一様乱数を使って求め、カプラン・マイヤー推定法による生存曲線の比較を行っていく。

7.2　結果・考察

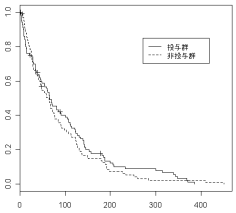

図 4: 2群の平均生存時間が同じ場合

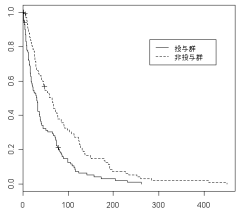

図 5: 2群の平均生存時間に差がある場合

2群の平均生存時間が同じ100の場合は、2つの生存曲線が同じような曲線を描いた。また、ログランク検定の結果から$p = 0.451$より有意水準5%とすると、2つの生存曲線との間に差があるとは言えないという結論に達した。次に2群の平均が違う場合は、平均の違いにもよるが、今回の結果では、生存曲線に明らかな違いが見られ、ログランク検定の結果から$p = 0.0000928$となり、有意であることが分かった。

8　おわりに

生存分析のデータは他のデータとは違い、打ち切りデータを考慮して解析をしなければならない。また、コックス比例ハザードモデルとカプラン・マイヤーでは、同じデータを用いて解析すると、共変量を導入しているコックス比例ハザードモデルの方がより深い解析ができることが分かった。

参考文献

[1] 岡田昌史: The R Book データ解析環境Rの活用事例集, (2004).

[2] 金明哲：フリーソフトによるデータ解析・マイニング第36回, Rと生存時間分析(1).

[3] Marcello Pagano,Kimberlee Gauvreau：生物統計学入門, 丸善株式会社 (2005).

[4] 高橋信：すぐ読める生存時間分析, 東京図書 (2007).

[5] Mantel, N, and Haenszel, W (1959) Statistical aspects of the analysis of data from retrospective studies of disease J Nat Cancer Inst 22,719-748.