mwp-086

生存時間分析の基本　【 評価版 】

Stata には生存時間解析用のコマンドが一式用意されています。機能区分という点で言うと、これらは

- ノンパラメトリック系機能
- セミパラメトリック系機能
- パラメトリック系機能

の 3 つに分類できます。本 whitepaper ではこれらの生存時間解析機能を使用して行く上で前提となる基本的な事項や概念について、情報を整理しておきます。

パネルデータモデル、混合効果モデル中での生存時間分析機能についてはそれぞれ [XT], [ME] ボリュームを参照ください。

1. 生存関数とハザード関数

今、ある事象が発生するまでの時間を X とします。その事象は短時間で発生することもあれば、なかなか発生しないこともあるでしょう。その意味で X は確率変数であるわけですが、その確率密度関数 (probability density function) を $f(x)$ で表すことにします[*1]。次の図は Weibull 分布

$$f(x) = \frac{px^{p-1}}{\lambda^p} \cdot \exp(-(x/\lambda)^p)$$

の場合を例に取って $f(x)$ の形状をプロットしたものです。

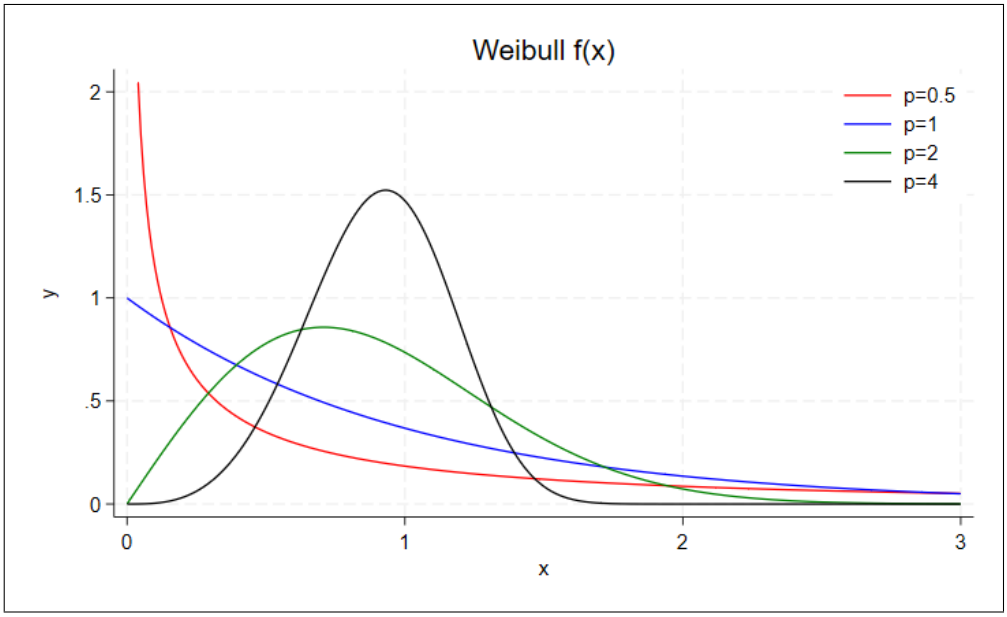

図 1　Weibull 分布の確率密度関数

Weibull 分布は形状パラメータ p の値によってさまざまな分布が表現できることから生存時間解析の分野では良く用いられる関数であり、$p = 1$ の場合の特殊ケースとして指数分布を中に包含しています。いずれにせよ確率密度関数 $f(x)$ によって x の分布が規定されるわけです。次に $f(x)$ が決まると累積分布関数 (cumulative distribution function) が

$$F(x) = \Pr(X \leq x) = \int_0^x f(t)dt \tag{1}$$

のように計算できます。

[*1] x は離散変数である場合も考えられますが、ここでは連続変数の場合に限って議論を進めることにします。

$f(x)$ も $F(x)$ も数学の分野ではきわめて一般的な概念を表しているわけですが、残念なことに生存時間解析の分野では余り使われません。代りとなるのが生存関数 (survivor function) であり、ハザード関数 (hazard function) であるわけです。生存関数 $S(x)$ は累積分布関数の補集合をなすものとして

$$S(x) = \Pr(X > x) = \int_x^{\infty} f(t)dt \tag{2}$$

のように定義されます。すなわち $S(x) = 1 - F(x)$ の関係にあるわけで、1 から 0 に向かって単調に減少して行く関数となります。事象が起きる時点が $X > x$ である確率を意味します。事象として死亡を考えるなら、時点 x よりも長生きする確率を表す関数と言えます。

一方、ハザード関数 $h(x)$ は条件付き確率 $\Pr(x \leq X < x + \Delta x | X \geq x)$ を使い、

$$h(x) = \lim_{\Delta x \to 0} \frac{\Pr(x \leq X < x + \Delta x | X \geq x)}{\Delta x} \tag{3}$$

のように定義されます。書き直すと

$$h(x) = \frac{1}{S(x)} \lim_{\Delta x \to 0} \frac{\Pr(x \leq X < x + \Delta x)}{\Delta x} = \frac{f(x)}{S(x)} \tag{4}$$

さらに $f(x) = -\frac{dS(x)}{dx}$ と書けることに注意すると

$$h(x) = \frac{f(x)}{S(x)} = -\frac{d}{dx} \ln[S(x)] \tag{5}$$

という関係式を導くことができます。

評価版では割愛しています。

2. データセットの初期設定

評価版では割愛しています。

3. ノンパラメトリック解析

ある特定の事象が発生するまでの時間 t を 1 組の共変量 x_k $(k = 1, 2, \ldots)$ との絡みで分析したいということであれば、それなりのモデル式の設定を伴うので、後述するセミパラメトリック解析、あるいはパラメトリック解析手法を用いる必要があります。しかし観察対象が全体的に一様である、あるいはグルーピングした場合には、それぞれのグループ内において観察対象が一様であり、共変量の影響は特に考えなくても良いといったケースも考えられます。そのような場合には何のモデル設定も行わず、単に観測された生存時間データから生存関数やハザード関数を推定するといった手法 — ノンパラメトリック解析 — を用いることができます。

3.1 Kaplan-Meier 生存関数

生存関数 $S(t)$ をノンパラメトリックな形で推定する際に一般的に用いられるのが Kaplan-Meier によって提唱された推定法です。対象とする事象（死亡や故障等）が時点 t_j $(j = 1, 2, \ldots)$ で発生したとします。それぞれの時点において生じた事象の数を d_j、そのときにリスク状態にある (at risk) 対象者 (subjects) の数を r_j としたとき、生存関数の推定値 $\widehat{S}(t)$ は

$$\widehat{S}(t) = \prod_{j(t_j \leq t)} \left(1 - \frac{d_j}{r_j}\right) \tag{6}$$

で与えられます。この一般式の形ではわかりにくいかも知れないので、簡単な具体例に従って見て行くことにします。特に途中打切り (censoring) の扱いについて注意してください。

3.2 Kaplan-Meier 生存関数の推定例

評価版では割愛しています。

3.3 比較検定

評価版では割愛しています。

4. セミパラメトリック解析

4.1 Cox 比例ハザードモデル

評価版では割愛しています。

4.2 部分尤度

評価版では割愛しています。

4.3 Cox 比例ハザードモデルの推定例

評価版では割愛しています。

4.4 層別化モデル

評価版では割愛しています。

5. パラメトリック解析

5.1 パラメータ化の流儀

評価版では割愛しています。

5.2 パラメトリックモデルの推定例

評価版では割愛しています。

補足１ – グラフ作成コマンド操作

評価版では割愛しています。

■