

『極値統計学』演習とデータ解析

高橋 倫也

平成28年 8月

概要

『極値統計学』の第2章～第4章の演習と第5章データ解析に関する補足説明。

1 はじめに

ソフトウェア R のパッケージ `ismev` は, Coles (2001) が彼の本 *An Introduction to Statistical Modeling of Extreme Values*. (ISMEV) でデータ解析に用いたソフトウェア S-Plus の関数を R で使えるようにしたものである。

『極値統計学』の演習とデータ解析では主に次の `ismev` の関数を使う：
ブロック最大モデルでは,

`gev.fit`, `gev.diag`, `gev.profxi`, `gev.prof`.

閾値超過モデルでは, `mrl.plot`, `gpd.fitrange`,

`gpd.fit`, `gpd.diag`, `gpd.profxi`, `gpd.prof`.

演習用に上記の関数の一部を変更したものなどが「演習用関数」にある。演習用関数の `cgev.diag` は `gev.diag` の PP-plot や QQ-plot に著者が手を入れた（見やすくした）ものである。また, `cgev.profxi`, `cgev.prof` は `gev.profxi`, `gev.prof` では信頼区間の下端と上端を求めてくれないので, それらを求めるようにした関数である。`gumQ` はグンベル確率紙へのプロットを描かず関数である。同様に, `cgpd.diag`, `cgpd.profxi`, `cgpd.prof` は著者が手を入れた関数。

以下, 『極値統計学』の各章の R による演習やデータ解析について説明して行く。Web サイトにある「演習とデータ解析」の該当するプログラムや「演習用関数」の関数などをコピーして, R の入力画面（ターミナル・ウインドウ）にペーストすれば演習やデータ解析が行える。

なお, R や `ismev` のバージョンによっては『極値統計学』の推定結果などと若干異なる値が求まることがある。

2 第2章の演習

この演習では、極値データにあてはめる分布として導出した、

1. 一般極値分布 $GEV(\mu, \sigma, \xi)$
2. 上位 r 個の同時漸近分布 $rGEV(\mu, \sigma, \xi)$
3. 一般パレート分布 $GP(\sigma, \xi)$

に従う乱数を生成し、それらのヒストグラムを描いて各分布の特徴を調べる。

「演習とデータ解析」の第2章演習のプログラムや関数を順次コピーし、Rの入力画面にペーストすれば演習が行える。

乱数生成プログラムの種 (seed, 整数), 生成数 $n = n$, パラメータ $\mu = m$, $\sigma = s$ (> 0), $\xi = xi$ などに種々の値を代入して分布の形状を調べてみよ。また, Gumbel 分布 $Gum(\mu, \sigma) = GEV(\mu, \sigma, 0)$ の乱数を生成して, そのヒストグラムを描いてみよ。

3 第3章の演習

ブロック最大モデルでは, `ismev` を用いてデータ解析を次のように行う。以下ブロック最大データを `data` とする。

1. `gumQ(data)`

まず, グンベル確率紙にブロック最大データをプロットし, その形状から GEV 分布をあてはめることができるかチェックする。

2. `a1 <- gev.fit(data)`

ブロック最大データに GEV 分布をあてはめて, 最尤推定値, 標準誤差などを求める。

3. `cgev.diag(a1)`

GEV 分布あてはめの診断図を描かす。診断図に問題がなければ次に進む。問題がある場合は, ブロックの大きさを大きくする。非定常のモデルを考えるなどの対策が必要。

4. `cgev.profxi(z, xlow, xup, conf = 0.95, nint = 500)`

形状パラメータ ξ の信頼係数 `conf = 0.95` の信頼区間を求める。 $z = a1$ とし, 区間 `[xlow, xup]` を信頼区間を含むように決める。 `nint` は区間 `[xlow, xup]` で対数プロファイル尤度を求める等間隔の点の個数で, 信頼区間の下端と上端は `nint` を大きくする (例えば, `nint = 500` または `1000`) と正確に求まる。以下にでてくる `z, xlow, xup, conf, nint` はここで紹介したものと同様のものである。

5. `cgev.prof(z = a1, m, xlow, xup, conf = 0.95, nint = 500)`

m 年再現レベルの 95% 信頼区間を求める。ここで, m は再現期間。以下にでてくる m も同じものである。

閾値超過モデルではデータ解析を次のように行う。ここでも, 観測データを `data` とする。

```
1. mrl.plot(data, umin = min(data), umax = max(data) - 0.1,  
   conf = 0.95, nint = 100)
```

data に対して、区間 [umin, umax] の等間隔の nint 個の点で平均超過と信頼係数 conf = 0.95 の信頼区間を求めてそれらをプロットする。この図からプロットが直線と見なせる閾値の候補範囲を見つける。

```
2. gpd.fitrangle(data, umin, umax, nint = 10)
```

上で見つけた閾値の候補範囲 [umin, umax] の等間隔の nint 個の点を閾値として、修正尺度と形状パラメータの推定値と 95%信頼区間を求め、閾値に対してそれらをプロットする。上の平均超過プロットとこの図から閾値 u を決定する。

```
3. a1 <- gpd.fit(xdat = data, threshold = u, npy = 365)
```

閾値を指定して GP 分布をあてはめて解析する。npy = 365 は年観測個数。

```
4. cgpd.diag(a1)
```

GP 分布あてはめの診断図を描かす。診断図に問題がなければ次に進む。問題がある場合は、閾値の再決定や非定常のモデルを考えるなどの対策をとる。

```
5. cgpd.profxi(z = a1, xlow, xup, conf = 0.95, nint = 500)
```

形状パラメータ ξ の 95%信頼区間を求める。

```
6. cgpd.prof(z = a1, m, xlow, xup, npy=365, conf=0.95, nint=500)
```

m 年再現レベルの 95%信頼区間を求める。

R を起動して、パッケージ ismev を読み込む。また「極値プログラム」にある関数をすべてコピーし R の入力画面にペーストする。次に、「演習とデータ解析」の第 3 章演習のプログラムを順次コピーし R の入力画面にペーストして演習を行う。

ここでの演習では関数 R5gev を読み込んでおく必要がある。

4 第 4 章の演習

ブロック最大モデルで、パラメータ (μ, σ, ξ) が時間 t に依存する非定常モデルの場合を考える。まず、データの時系列プロットとグンベル確率紙へのプロットを作製して、定常性をチェックする。定常性が疑われるので、非定常モデルでデータの解析を行う。

「演習とデータ解析」の第 4 章演習のプログラムを順次コピーし R の入力画面にペーストして演習を行なう。ここでは、パッケージ ismev と関数 cgev.diag を読み込んでおく必要がある。

5 第 5 章データ解析

第 5 章のデータ解析の一部を紹介する。R を起動して、パッケージ ismev を読み込む。また「演習用関数」のすべての関数をコピーし R の入力画面にペーストする。

ここで用いるデータは次の三つである。

DATA.txt 神戸の年最大日降水量 KobeM と測器補正した東京の年最大風速
cTokyow がある。これらをコピーして R の入力画面にペーストする。

Kobe0.txt 神戸の日降水量。このデータは R に読み込む必要がある。

まず、神戸の年最大日降水量 KobeM をグンベル確率紙にプロットしてその形状をみる。GEV モデルがあてはまりそうである。「演習とデータ解析」の第 5 章 神戸年最大日降水量の箇所を順次コピーして、R の入力画面にペーストしてデータ解析を実行する。

次に、神戸の日降水量 Kobe0.txt を R に読み込み、これに閾値超過モデルをあてはめて解析する。「演習とデータ解析」の第 5 章 神戸日降水量の箇所を順次コピーして、R の画面にペーストしてデータ解析を行う。

最後に、測器補正をした東京の年最大風速 cTokyow に非定常 GEV モデルをあてはめて解析する。「演習とデータ解析」の第 5 章 東京の年最大風速の箇所を順次コピーして、R の画面にペーストしてデータ解析を確認されたい。