土木学会論文集B2（海岸工学）
Vol. 66，No.1，2010，141-145

気象指数を共変量とする水域外力の極値モデルに対する外れ値の感度分析

Sensivity for Outliers of Sea Extremes with the Time and Climate Index Covariates

北野利一[1]・喜岡　涉[2]・高橋倫也[3]

Toshikazu KITANO, Wataru KIOKA and Rinya TAKAHASHI

Sea extremes (annual maximum sea levels, significant wave heights over a certain threshold, etc) will be modelled with a temporal trend, and they may be also governed by the climate factors, e.g. Southern Oscillation Index (SOI). The fitting becomes better in general when any explanatory variable is added in the regression model. The sensitivity for the residuals should be examined to avoid the over-fitting. The outliers detection for extreme values can be firstly discussed by the degree of experience, which we proposed in the previous study. It will conduct to the robustness of estimation. The judgements for the removal of outliers are demonstrated in a diagram of leverage and residual of extremes.

1. まえがき

波高や潮位などの自然外力の極値の傾向が変化しつつある可能性は，近年注視されている．その経年変化に加え，南方振動指数（SOI）や北極振動（AOI）などのより具体的な気象指数も変数として組み込めば，モデルへの当てはまりは良くなることは予想できる．しかしながら，過剰なフィッティングを避けるために，残差解析が必要となる．残差は，純粋な確率変動量と共変量の変動性に伴う量の2つに影響を受ける．特に，時間変数と気象指数のように複数の共変量を用いる場合には，単純なデータの図示による直感的な考察をするだけでは判断を誤る可能性が非常に大きい．

従来の極値解析法では，母数推定の後，モデルへの適合診断が検討されるのみである．本研究では，複数の共変量をもつ極値モデルからの外れ値を検出し，モデルの母数推定に際して，その外れ値を棄却すべきか（あるいは，容認してよいか）という判断を与える感度分析を導入するものである．このような検討は，希少頻度で生起するメガリスクの検討にこそ必要とされるロバスト（頑健）性の議論に不可欠と考える．

2. 複数の共変量を伴う極値モデル

図-1(a)は，年最大潮位（Fremantle港，豪州西岸）の時系列である．1901-1979年の79年間のデータを用いる（ただし，欠測のため，標本サイズは72）．図中の直線に示すような経年トレンドが想起される．また，**図-1**(b)は，同じ年最大潮位資料を南方振動指数（SOI）に対して表示しており，SOIが正の場合，年最大潮位が大きく，負の場合に潮位は低いという傾向が見られる．このような場合，時刻 x_1 (= ($\tilde{x}_1$ − 1940)/39，西暦 $\tilde{x}_1$年）とSOIの値 x_2が共変量となりうる．このデータをもとに，Coles(2001)は，非定常な極値モデルの構築法とモデル選択の議論をしている．

レベルyを超える外力の来襲する単位時間あたりの頻度を表すため，次式に示す生起強度が用いられる．

$$\lambda(y;\boldsymbol{\theta}) = \left(1+\xi\frac{y-\mu}{\sigma}\right)^{-1/\xi} \quad \cdots\cdots(1)$$

ここで，定数$\boldsymbol{\theta}$=(μ，σ，ξ) をそれぞれ年最大潮位分布の位置，尺度および形状母数とすれば，式(1)は，レベルyを超える潮位の年平均回数を表す．$\lambda(\mu;\boldsymbol{\theta})$=1となるので，潮位$\mu$は，年平均1回を超える潮位レベルに相当する．その潮位μを共変量x_jに関連づけて，

$$\mu = \mu_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \quad \cdots\cdots(2)$$

とモデル化できる．第4項目は，Coles(2001)では検討されていない交互作用項である．**図-1**(c)は，79年間を重なりを含めて6期間に分割し，各々の期間における潮位とSOIの関係を示したものである．後半3期間（上段）では，潮位とSOIの関係が見られるのに対し，前半3期間（下段）では，潮位とSOIの関係が明確ではない．これは，交互作用項の有無により検討できる．さらに，尺度母数や形状母数も同様に，共変量を関連づけたモデル化も可能である．ここでは，

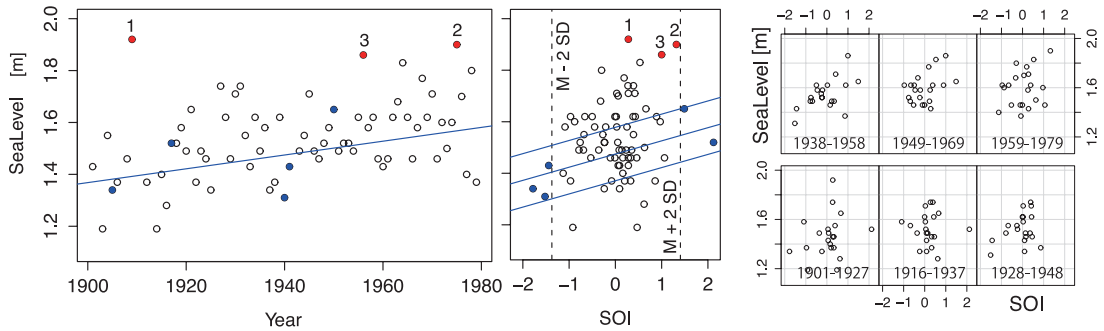
$$\log\sigma = \log\sigma_0 + \alpha_1 x_1 \quad \cdots\cdots(3)$$

$$\xi = \xi_0 \quad \cdots\cdots(4)$$

とする．式(1)で生起強度が与えられる点過程モデルとして，最尤法により母数を推定できる．モデルの選択にあたっては，**図-2**に示すとおり，AICを最小化する最適モデル（M4）の係数は，次のとおりである．

$$\alpha_1 = \beta_3 = 0;$$
$$\hat{\mu}_0 = 1.47,\ \hat{\beta}_1 = 0.1037,\ \hat{\beta}_2 = 0.0511, \quad \cdots\cdots(5)$$
$$\hat{\sigma}_0 = 0.124,\ \hat{\xi}_0 = -0.15$$

1 正会員　博(工)　名古屋工業大学大学院准教授　工学研究科
2 フェロー　Ph.D　名古屋工業大学大学院教授　工学研究科
3　工(博)　神戸大学大学院教授　海事科学研究科



（a）時系列表示　　（b）気象指数による表示　　（c）共変量の分割による表示

図-1　豪州西岸の潮位データ（既往最大1~3位のデータに番号を付し，SOIの絶対値が大きい点も塗りつぶしてある）

モデルへの適合診断の1つとして，QQプロットが用いられる．モデルの分位点（quantile）に対して，バラツキを伴うものの，データがほぼ一直線上に並んでいることを期待するものである．共変量を条件とする年最大潮位y_iが一般化極値分布に従う時，生起強度$\lambda(y_i;\boldsymbol{\theta})$が標準指数分布に従う．そのため，年最大潮位を降順に並び替えた順序統計量$y_{(i)}$に対して，次式で表される点を図示することにより，QQプロットが描ける．

$$\left\{-\log\left(1-\frac{i}{n+1}\right),\lambda(y_{(i)};\hat{\boldsymbol{\theta}})\right\} \quad \cdots\cdots(6)$$

なお，Coles(2001)では，$-\log\lambda(y_i;\boldsymbol{\theta})$が標準ガンベル分布に従うことを利用して，QQプロットを描いている．**図-3**を見るとおり，モデルへの適合が良好である．他の適合診断法として，次式の点を図示することによる PPプロットも知られる（図示は割愛する）．

$$\left\{\frac{j}{n+1},\ \exp\left[-\lambda(y_{(n-j)};\hat{\boldsymbol{\theta}})\right]\right\} \quad \cdots\cdots(7)$$

年平均1回を超えるレベルμ_iに対して，$1/\lambda(y_i;\hat{\boldsymbol{\theta}})$の確率変動（後に示すとおり，これは再現期間に相当する）が均質であることも診断の1つとなる（図-4参照）．

Coles(2001)およびBeirlantら(2004)をはじめとして，極値解析手法を論じた既往の研究では，最適モデルの採択後，QQおよびPPプロットによる適合診断で終始しており，個々のデータがモデルに及ぼす影響を検討する感度分析は，これまで一切議論されていないのが現状である．感度分析の観点から考えれば，例えば，図-1(b)にて，SOIが大きい（小さい）値となる時，平年値を下（上）回る年最大潮位が出現していない現状が偶然によるものかもしれない．SOIの絶対値が大きくなる頻度が少ないにも関わらず，得られる推定結果を信頼できるか否かを判断することである．

3. 共変量を伴う極値に対する経験度

(1) 生起強度と経験度

あるレベルを超える外力の平均生起間隔を再現期間と定義するなら，再現期間R年の確率外力y_Rは，次式の関係を満たす．

$$\lambda(y_R;\boldsymbol{\theta})=\frac{1}{R} \quad \cdots\cdots(8)$$

この時，ポアソン分布と共役となるガンマ分布の形状母数が，生起強度の推定量$\hat{\lambda}(=\lambda(y_R;\hat{\boldsymbol{\theta}}))$を算出する際の実質的な標本サイズに相当することを利用して，北野ら(2008)は経験度という概念を導入した．その結果，生起強度の変動係数Cを用いて，経験度Kは

$$K=\frac{1}{C^2};\quad C=\frac{\sqrt{V(\hat{\lambda})}}{E(\hat{\lambda})} \quad \cdots\cdots(9)$$

と与えられる．このように非常に単純な表現となることも，経験度の特徴の1つと言える．経験度を具体的に算出するため，デルタ法により，生起強度の誤差分散を母数の誤差分散$V_{\boldsymbol{\theta}}$で変換して，

$$K(y_R)=\frac{\lambda^2(y_R;\boldsymbol{\theta})}{\nabla'\lambda(y_R;\boldsymbol{\theta})\,V_{\boldsymbol{\theta}}\,\nabla\lambda(y_R;\boldsymbol{\theta})} \quad \cdots\cdots(10)$$

と変形する．ここで，∇は母数による微分を表し，∇'はその転置ベクトルである．

(2) 共変量による母数が変化する場合の経験度

式(5)で採択されたモデル，すなわち，位置母数のみを共変量でモデル化する場合には，母数とその微分は，

$$\boldsymbol{\theta}=(\mu_0,\beta_j,\sigma_0,\xi_0) \quad \cdots\cdots(11)$$

$$\nabla'=\frac{\partial}{\partial\boldsymbol{\theta}}=\left(\frac{\partial}{\partial\mu_0},\frac{\partial}{\partial\beta_j},\frac{\partial}{\partial\sigma_0},\frac{\partial}{\partial\xi_0}\right) \quad \cdots\cdots(12)$$

となる．この時，共変量に係る母数β_jを除いた母数ベクトルおよび微分ベクトルをそれぞれ，$\boldsymbol{\theta}_0$および∇_0と記す．また，母数の誤差分散は，フィッシャーの期待情報行列から与えられるものとして，$V_{\boldsymbol{\theta}}$ および $V_{\boldsymbol{\theta}_0}$と記す（後者は，母数$\boldsymbol{\theta}_0$のみの誤差分散行列）．

式(10)の$\boldsymbol{\theta}$に$\boldsymbol{\theta}_0$を形式的に代入して得られる経験度（これを狭義の経験度とよぶ）は，

$$\frac{1}{K_0}=\left.\frac{\nabla_0'\lambda\,V_{\boldsymbol{\theta}_0}\,\nabla_0\lambda}{\lambda^2}\right|_{\boldsymbol{\theta}_0}=q(\lambda)+\frac{1}{N} \quad \cdots\cdots(13)$$

となる．$q(\lambda)$は，生起頻度λのみならず，形状母数ξ_0および標本サイズN（この場合，年数）にも依存する．例

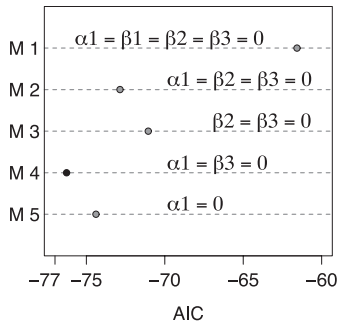


図-2　モデル選択

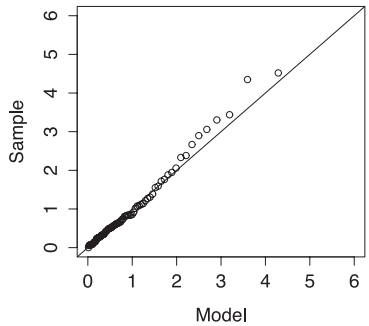


図-3　確率変動の適合性

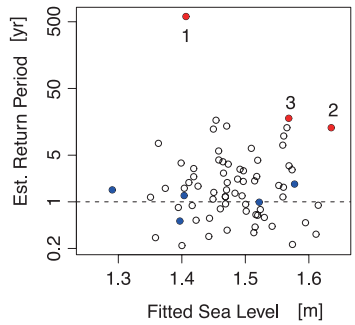


図-4　確率変動の均質性

えば，ガンベル分布（$\xi_0 = 0$）に従う年最大値データを扱う場合であれば，次式のように得られる．

$$q(\lambda) = \frac{6}{N}\left(\frac{1-\gamma-\log\lambda}{\pi}\right)^2 \quad \cdots\cdots(14)$$

式(13)で与えられる経験度K_0は，共変量による変化を伴わない．したがって，モデルに対する残差成分，この場合，極値分布による（純粋な）確率変動である．

次に，式(10)を変形すれば，次式が一般的に成り立つ．

$$\frac{1}{K} = D^2(\boldsymbol{x}_i) + \frac{1}{K_0} \quad \cdots\cdots(15)$$

ここで，平均値ベクトル$\bar{x}$および行列S

$$\bar{x}_i = \sum_k x_{ki} \,/\, N \quad \cdots\cdots(16)$$

$$S_{ij} = \sum_k (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j) \quad \cdots\cdots(17)$$

を用いて，次式で表される量を用いている．

$$D^2(\boldsymbol{x}_i) = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})' \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \quad \cdots\cdots(18)$$

判別分析では，集団の中心 $\bar{x}$ からデータx_Iの乖離量を

$$MD_i = \sqrt{N-1}\, D(\boldsymbol{x}_i) \quad \cdots\cdots(19)$$

と表されるマハラノビス距離で計る．さらに，回帰分析において，個々のデータが回帰モデルに及ぼす影響を検討する際に，次式で表されるてこ比が用いられる．

$$h_{ii} = D^2(\boldsymbol{x}_i) + \frac{1}{N} \quad \cdots\cdots(20)$$

なお，てこ比がとりうる範囲は［1/N, 1］である．

式(13), (15)および(20)の表現の類似性から，経験度は，てこ比の拡張であり，標本サイズに類するものであることが確認できる．さらに，以下のように整理できる．

$$\frac{1}{K} + \frac{1}{N} = \frac{1}{K_0} + h_{ii} \quad \cdots\cdots(21)$$

これにより，共変量を伴う極値モデルに対する経験度は，確率変動による狭義の経験度K_0と共変量による変動量を表すてこ比h_{ii}に分解できることが明確になる．

(3) 回帰分析における感度分析との関係

この節で扱う回帰モデルとは，pを母数の数として，

$$E(y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} \quad \cdots\cdots(22)$$

で表される．このような回帰モデルにデータを当てはめる際には，母数推定に影響を与える外れ値を検出し，外れ値を除去した上で，改めて母数推定をすべきである．推定された回帰直線から個々のデータが乖離していることは，残差から判断できる．ただし，推定される残差eは一様ではないので，基準化された残差rを用いる必要がある．一方，外れ値が回帰直線を引っ張って，見当違いの母数推定を行っているような場合には，外れ値の残差は，それほど大きくならない．そのため，残差のみによる外れ値の検出は良くないので，Cook(1977)は，外れ値の候補となるデータx_iを除いて得られる推定量$\hat{y}_{(i)}$に対して，除かずに（全てのデータを用いて）得られる推定量$\hat{y}$との差分を検討することにより，感度分析ができることを示している．すなわち，

$$CD_i = \frac{(\hat{\boldsymbol{y}}_{(i)} - \hat{\boldsymbol{y}})'(\hat{\boldsymbol{y}}_{(i)} - \hat{\boldsymbol{y}})/p}{\boldsymbol{e}'\boldsymbol{e}/(N-p)} \quad \cdots\cdots(23)$$

を外れ値を検出する指標として提案している．さらに，

$$r_i = \frac{e_i}{\sqrt{1-h_{ii}}} \Big/ \sqrt{\frac{\boldsymbol{e}'\boldsymbol{e}}{N-p}} \quad \cdots\cdots(24)$$

と表される規準化残差を用いれば，式(23)のクックの距離を次式のように書き換えることができる．

$$CD_i = \frac{r_i^2}{p}\frac{h_{ii}}{1-h_{ii}} \quad \cdots\cdots(25)$$

以上の結果として，推定された回帰モデルに対する規準化された残差r_i^2と共変量によるてこ比h_{ii}の2項に分解されることがわかる．単純な例として，ほ乳類の脳の重さzを，体重wで説明する回帰モデル

$$E(\log z\,|w) = \beta_0 + \beta_1 \log w \quad \cdots\cdots(26)$$

を考える．図-5には，てこ比h_{ii}を横軸に，規準化された残差r_i^2を縦軸にとり，クックの距離CD_iの等高線図を描いたものである．回帰モデルに対して，このような図示法はRLプロットとよばれている（渋谷・柴田，1992）．てこ比の大きなアフリカ象やアジア象は，共変量（この場合，体重）が大きく外れているため，回帰係数の推定に対する影響力が大きく，ヒトやオポッサムは，モデルに対する当てはまりが悪い（残差が大きい）．クックの

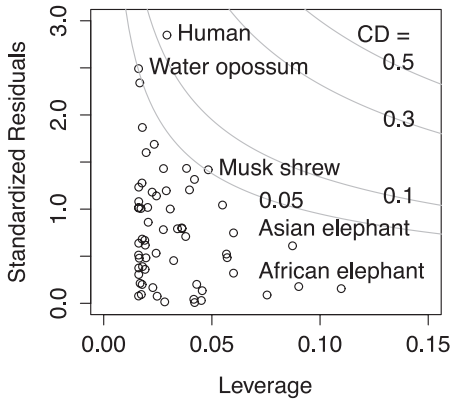


図-5　てこ比と残差に分解されるクックの距離

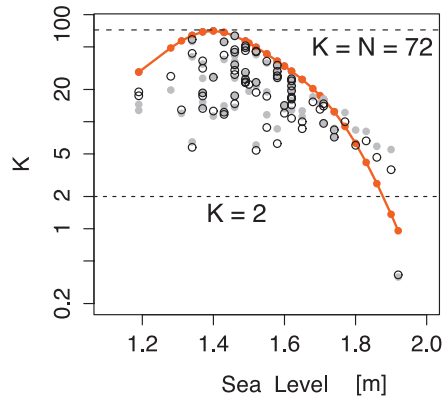


図-6　潮位データ（Fremantle 港）に対する経験度の算出

距離が大きな値（$> F_{p,N-p,0.5} \approx 1$）をとる外れ値は，除去されるべきと考える（Gross，2003）が，この場合は，該当する外れ値は無い．

極値モデルに対する感度分析も，あるデータを除いて得られる推定値と除かずに（全てのデータを用いて）得られる推定量の差分を考えるべきであろう．しかし，正規モデルとは異なり，極値モデルに対して，そのような正攻法が功を奏するとは考えにくい．むしろ，式(21)を見るとおり，既に，極値モデルに対する確率変動量（残差）K_0と共変量によるてこ比h_{ii}の2項に，経験度が分解されている．これは，残差とてこ比の2項に，クックの距離が分解されることに対応している．その結果として，極値解析の外れ値に対する感度分析に，式(21)を応用しようという考えである．ただし，式(21)は，観測データから推定された量から導かれた関係式ではなく，母数の誤差分散は，真値まわりに展開した理論的に得られるフィッシャーの期待情報行列から得たものである．すなわち，有限のデータの総和から求めたものではなく，数学的な極限操作である積分により求めている．観測データに対しては，最尤推定の計算過程で得られる観測情報行列を用いて，母数の誤差分散$\hat{V}_\theta$および$\hat{V}_{\theta_0}$を求め，式(10)のように変換すれば，観測潮位y_iに対する経験度 $K(y_i)$および狭義の経験度 $K_0(y_i)$を算出することができる．しかし，残念ながら，極値分布が指数分布族でないことから，式(21)に示される関係は厳密には成立せず，近似的になることに注意を要する．

図-6では，経験度 $K(y_i)$（白丸）に対して，狭義の経験度 $K_0(y_i)$とてこ比から，式(21)を用いて近似的に算出したもの（薄灰色の点）である．また，線分で繋いだ点は，定常モデルで当てはめて算出した経験度である．共変量を考慮したモデルでは，同じ潮位でも確率変動量が同じではないので，経験度は異なる値をとる．

4. 年最大潮位モデルに対する外れ値の感度分析

図-7は，経年および気象指標による変動量を表すてこ比h_{ii}の逆数を縦軸に，極値変動のみによる狭義の経験度$K_0(y_i)$を横軸にとり，観測データをプロットしている．なお，本研究の興味の対象となる経験度 $K(y_i)$には，その値を添えている．また，式(21)で表される（理論値として得られる）経験度Kの等高線も描いており，経験度$K(y_i)$の値は等高線の示す値から若干異なるが，近似は良好といえる．

図-8は，潮位に対する狭義の経験度およびてこ比を示している．図-7だけでは，潮位との関係が不明になるので，このような補助図が必要となる．また，図-7および図-8において，塗りつぶしてある点は，潮位が既往最大1～3位の点と，SOIの値が平均値から標準偏差の2倍以上に乖離する点を示しており，これらの点を外れ値の候補と考え，以下に議論する（図-1も参照，それらの点は，同様に塗りつぶしてある）．

経験度が $K<2$となるデータを外れ値と考える．ただし，無条件に，外れ値を除外すべき異常値と考えてはいけない．条件によって，2つの選択肢がある．北野ら(2008)で述べるとおり，経験度は，推定に係る実質的な標本サイズを意味するので，その限界値を2としている．狭義の経験度が $K_0<2$となるため，経験度も $K<2$となっている場合は，極値変動が大きく乖離している（そのデータに対して推定される再現期間は長い）ため，その推定値を信頼できないと考える．すなわち，その外れ値に対する必要な情報が不足しており，判断を保留する．したがって，モデルの母数を推定する際には，その外れ値も含める．しかし，その母数推定の結果を用いて，その外れ値に対して推定される内容を保留するのである．もちろん，外れ値以外の（経験度が十分に大きな）データに関する推定結果は利用できる．その一方で，てこ比が大きいために，経験度が $K<2$となるものは，共変量

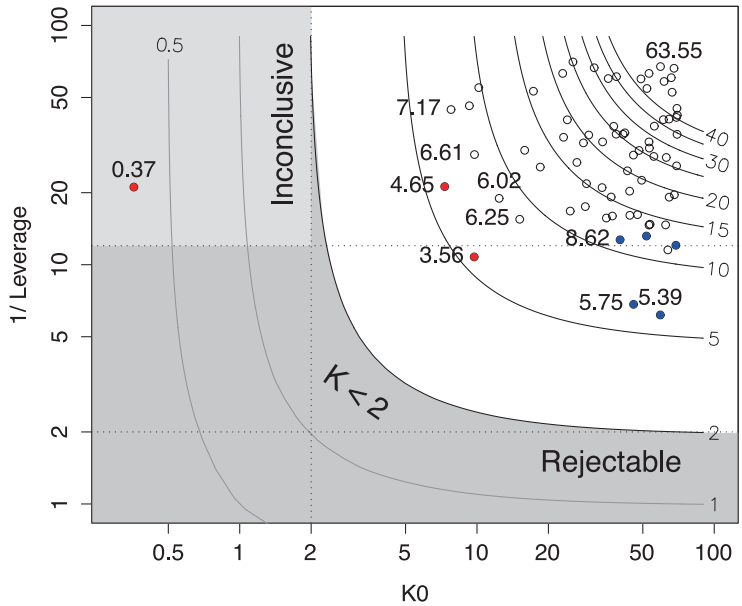


図-7　てこ比と残差に分解される経験度

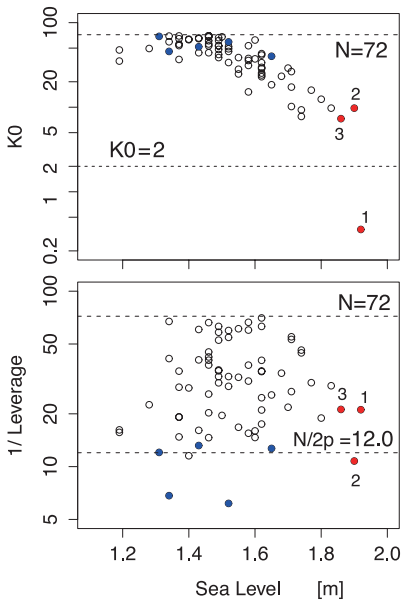


図-8　てこ比と残差に対する潮位の関係

が外れ値となり，乖離していることが原因である．そのため，共変量が同等な値をとる観測数を増やすことができれば，てこ比を減少できる．しかし，自然外力の観測は，“共変量を制御できる実験”ではないので，それは困難であろう．現実的には，このような外れ値を除外して，改めて母数推定を行うべきである．

図-7および図-8（下）を見るとおり，SOIの値が乖離しているデータは，てこ比が大きい．経験度と同じように，てこ比の逆数の限界値を2（これは，てこ比の限界値を0.5とする Huber(1981)の条件に相当）とすると，どのデータも外れ値とはならない．しかし，てこ比の限界値を$2p/N$とする条件（例えば，Gross(2003)を参照）を採用すれば，SOIの絶対値が大きな2つのデータと潮位が既往第2位のデータのてこ比が $1/h_{ii}<12.0$となる．しかしながら，経験度は $K>2$であるので，除去すべき外れ値とはならない．したがって，SOIの絶対値が大きいデータは疎であるが，SOIの変化による潮位の変化の傾向を求めるために使用してよいことがわかる．また，潮位の既往最大値の経験度は $K<2$であるが，てこ比はそれ程大きくない（てこ比の逆数は小さくない）．潮位の既往最大値も，除去すべき外れ値とはならない．しかし，狭義の経験度は $K_0<2$となるので，このデータに対する推定結果（例えば，図-4を見るとおり，100年にも満たない観測期間に対して，既往最大潮位の再現期間は500年を超えている）についての判断を保留する．すなわち，既往最大値の再現期間は推定不能と判断せざるを得ない．しかし，モデルの母数推定には，既往最大値を使用する（除外しない）．

5. まとめ

現時点の観測資料に，水域外力の極値モデルが過剰に適合することは，将来の観測データに対してロバストでない．できる限り長期的に堅牢な防災施設計画を行う基礎資料として，経験度を応用すれば，推定結果に対する個々のデータの感度分析も可能であることを示した．

謝辞：本研究は，科学研究費（代表：北野利一，課題名：経験度による推定可否判断を付した水域メガリスク解析，代表：高橋倫也，課題名：極値理論によるリスクの予測）の助成による成果の一部である．また，ICHARMとの共同研究（課題名：非定常過程における水文頻度解析）の成果の一部である．

参　考　文　献

北野利一・森瀬喬士・喜岡　渉・高橋倫也（2008）：確率波高に対する推定の可否を決定づける新たな指標の提案，海岸工学論文集，第55巻，pp.141-145.

渋谷政昭・柴田里程(1992)：Sによるデータ解析，共立出版，228p.

Beirlant, J., Goegebeur, Y., Segers, J., and J. Teugels (2004): Statistics of extremes: theory and applications, Wiley, 490p.

Coles, S. (2001): An introduction to statistical modeling of extreme values, Springer, 208p.

Cook, (1977): Detection of influential observations in linear regression, Technometrics, Vol. 19, pp.15-18.

Gross, J. (2003): Linear regression, Lecture Notes in Statistics, 175, Springer, 394p.

Huber, P. J. (1981): Robust statistics, Wiley, 308p.