OUTLIER SENSITIVITY ON THE SEA EXTREMES BY THE TEMPORAL AND CLIMATE INDEX COVARIATIONS

Toshikazu Kitano[1], Wataru Kioka[1] and Rinya Takahashi[2]

Outlier detection is one of the classical problem in the regression analysis. For this purpose the Cook's distance was proposed as the amount of changing the predictions by removing the candidate outlier in comparison with the total variation of the residuals against the fitting plane. This distance is considered to be so useful that it is rearranged and discribed in the two terms of the leverage of covariates and the contingent discrepancy. Hence the outlier detection can be displayed as a diagram with these two terms. Extremes generally accompanies outliers. Unfortunately the Cook's distance wouldn't be applicable to the outlier among the extremes. It is one of the reason that the extreme value distribution doesn't belong to the exponential family. Thus we should find the alternative way. The degree of experience, proposed originally for evaluating the limitation of extrapolation, will play an important role of detecting the outliers, because it is decomposed into two parts of the leverage of covariates and the contingent discrepancy in the average sense. Not only the mathematical derivations are shown but also a practical judgement for the removal of outliers is demonstrated in a diagram of leverage and residual of extremes.

Keywords: degree of experience; influential outlier; return period; climate index

Introduction

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 is extended by adding the leverage term to those proposed in the previous study shown in Kitano et al. (2008, 2009, 2010, 2011). It will conduct to the robustness of estimation.

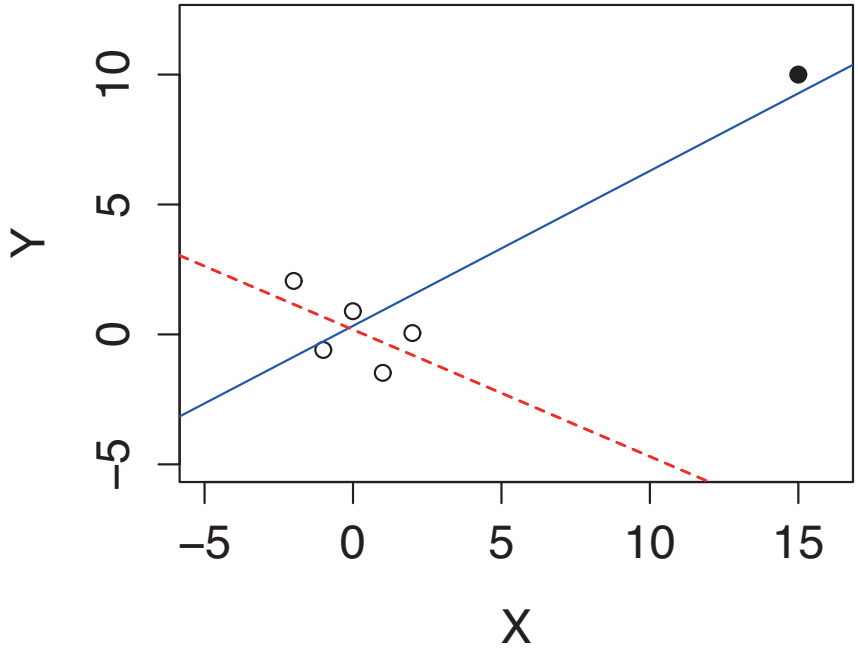
We have an essential problems in the statistical analysis to evaluate the return levels of sea extremes for the design of coastal structures, which is due to the poverty of the available data. It should be called the small sample size problems, and they bring some practical questions to us in the following two points of view: 1) Limiation of extrapolation, and 2) Sensitivity against outliers.

The point 1) is a problem arised when the resultant statistical model is applied, while the point 2) is one arised when an examining statistical model is fitted. As seen in Kitano et al. (2008), the degree of experience is proposed for the limitation of the quantile extrapolation. It is a simplest extrapolation, in which the return levels are obtained by extending the fitted quantile line against the data set regarding as beeing extracted from an identical population. Kitano et al. (2010) modified the concept of the degree of experience to be applied to a non-stationary models (with a temporal trend), and demonstrated the limitation of the temporal extrapolation as well as the return levels with the confidence intervals for the sea level of Venice. On the basis of the uncertainty accumpanied with a trend, Kitano et al. (2010) pointed out that the uncertainty increases against the passage of time even for the stationary model, and it is named the diffractive effect.

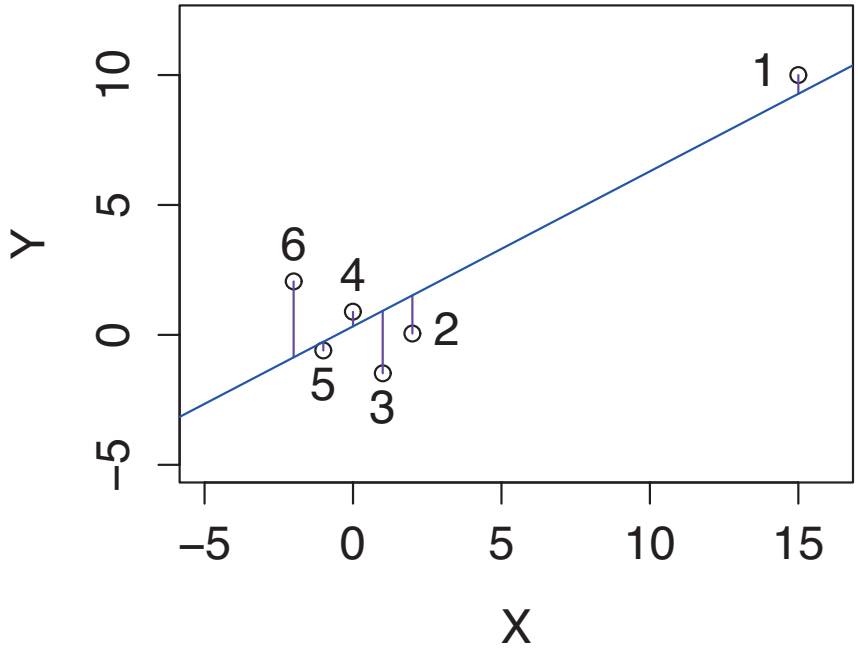
In these studies, the degree of experience is used as the post-analysis after the target model is fitted to the observed data, as mentioned before. As the pre-analysis, during a model is tested to be fitted, we sometime face to an influencial data, which pulls the model near oneself, and we bothered if the data should be removal or not. This is known as the sensitivity analysis in the regression analysis, where the response variates conditionally with the covariates. The set of covariates has not always concentrated but it has also some periphery parts, where the data is so poor to lead to a kind of small sample size problem. It is optimistic that sea extremes are always considered to be extracting the identical population. In some cases sea extremes will be covariated with the climate index, for example, SOI, AOI, and the average sea surface temperature, etc. Therefore, the sensitivity against outlier should be discussed for sea extremes.

[1] Dept. of Civil Engineering, Nagoya Institute of Technology, Gokisocho, Showaku, Nagoya, 466-8555, Japan

[2] Faculty of Maritime Science, Kobe Univ. Fukue-minamimachi, Higashinadaku, Kobe, 658-0022, Japan



a) Two fitting lines before and after removing the outlier



b) Residuals helpless for detecting the outlier

Figure 1 Outlier in the regression analysis

A Treatment of Outlier in Regression Analysis

Here we reconfirm the treatment of outlier in the general regression analysis as the common knowledge. We here consider a statistical model as the following:

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

where we take multi-covariates x_i (the number of covariates is $p-1$, and including an intercept, the number of the parameters is p) in general sense, and we can reduce it to a single covariate easily at

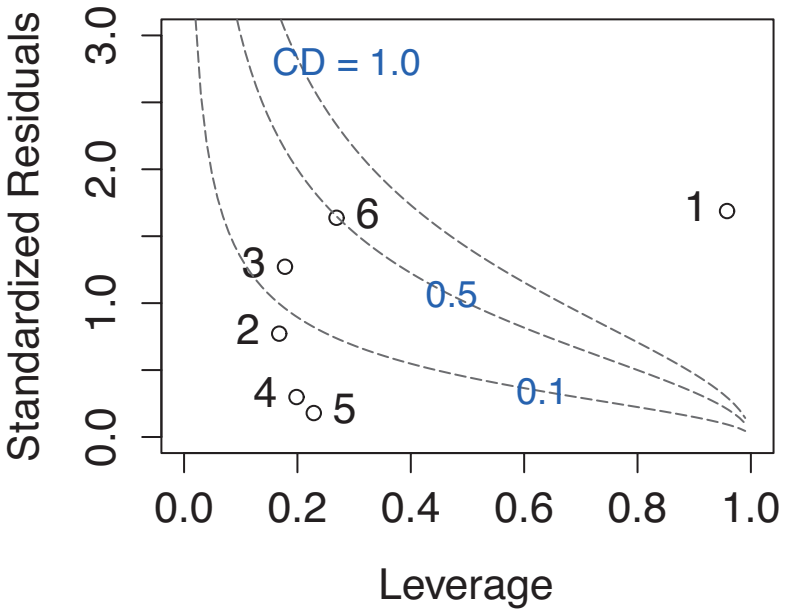


Figure 2 Outlier sensitivity diagram by means of the Cook's distance in regression analysis

any stage of the following procedure. As mentioned before, Cook (1977) introduced an index for outiler by the amout of difference between the estimation $\hat{\boldsymbol{y}}$ by all data and the one $\hat{\boldsymbol{y}}_{(i)}$ by the data removing the target data (x_i, y_i) compared with the amount of the residual variation $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$, which is named Cook's distance difined by

$$CD_i = \frac{\left(\hat{\boldsymbol{y}}_{(i)} - \hat{\boldsymbol{y}}\right)' \left(\hat{\boldsymbol{y}}_{(i)} - \hat{\boldsymbol{y}}\right)/p}{\boldsymbol{e}'\boldsymbol{e}/(n-p)} \tag{2}$$

It should be noted that the residual errors depend on the covariates' values. Therefore, we use the standarized residual defined as the followings:

$$r_i = \frac{e_i}{\sqrt{1-h_{ii}}} \Bigg/ \sqrt{\frac{\boldsymbol{e}'\boldsymbol{e}}{n-p}} \tag{3}$$

As consequence, we transformed the Cook's distance into the form with the statistical variation and the leverage:

$$CD_i = \frac{r_i^2}{p} \Bigg/ \left(\frac{1}{h_{ii}} - 1\right) \tag{4}$$

where is the leverage of the target covariate h_{ii}, and the detail definition will be shown later for the multicovariate case. According to the range of leverage is

$$\frac{1}{n} \leq h_{ii} < 1 \tag{5}$$

it is found that the Cook's distance becomes larger in the case that the statistical variation is larger, or the case that the leverage is large and close to 1, or both cases. An index is not only defined but also transformed in the interpretable expression in the point of view of knowing clearly how to work.

Fig. 2 is shown the contourlines of the Cook's distance against the normalized residuals and the leverage, and the data named as 1 is clear to be an outlier whose leverage is very high though the residual is not so large. Therefore, we judge that it should be removable as an outlier due to high-leverage. The diagram as shown in Fig. 2 is very useful and indispensable for the outlier judgement. But it was invented for the ordinal regression analysis, it isn't easily applicable to the extremes. We should make another invention for extremes, and we can think that also for this purpose the degree of experience proposed by Kitano et al.(2008) will works comprehensively in the place of Eq.(2).

Degree of experience including covariates

As the effective size number of the sample in the contribution to estimate the extrapolating level, by considering the Fisher's information against the occurrence rate and more interpreting it the shape parameter's value of a natural conjugate gamma distribution in the point of view of Bayesian inference, Kitano et al. (2008) proposed the degree of experience K given by

$$\frac{1}{K} = V(\log\lambda) \tag{6}$$

where the occurence rate is defined as

$$\lambda(y) = \exp\left\{-\frac{1}{\xi}\log\left(1+\xi\,\frac{y-\mu}{\sigma}\right)\right\} \tag{7}$$

in terms of the location, scale and shape parameters $\boldsymbol{\theta} = \{\mu, \sigma, \xi\}$ of the generalized extreme value distribution (GEV). Especially in case of Gumbel type $\xi = 0$, Eq.(7) becomes the following simple function.

$$\lambda(y) = \exp\left(-\frac{y-\mu}{\sigma}\right) \quad \text{for} \quad \xi = 0 \tag{8}$$

Since the deviation of $\log\lambda$ becomes, like the derivative,

$$\delta\log\lambda = \frac{\delta\lambda}{\lambda} \tag{9}$$

the degree of experience can be transformed into the following amount:

$$\frac{1}{K} = \frac{V(\lambda)}{\{E(\lambda)\}^2} \tag{10}$$

This is corresponding with the properties of the natural conjugate gamma distribution for a Poisson distribution including the occurrence rate of Eq.(7). The gamma distribution with the parameters of a shape parameter K and an effective time length of observation L, described by

$$f(\lambda) = \frac{(L\lambda)^K}{\lambda\Gamma(K)}\,e^{-L\lambda} \tag{11}$$

are shown in Fig.3, and they are concentrated around the mean occurence

$$E(\lambda) = \frac{K}{L} \tag{12}$$

and it is governed by values of the shape parameter K. The variance is

$$V(\lambda) = \frac{K}{L^2} \tag{13}$$

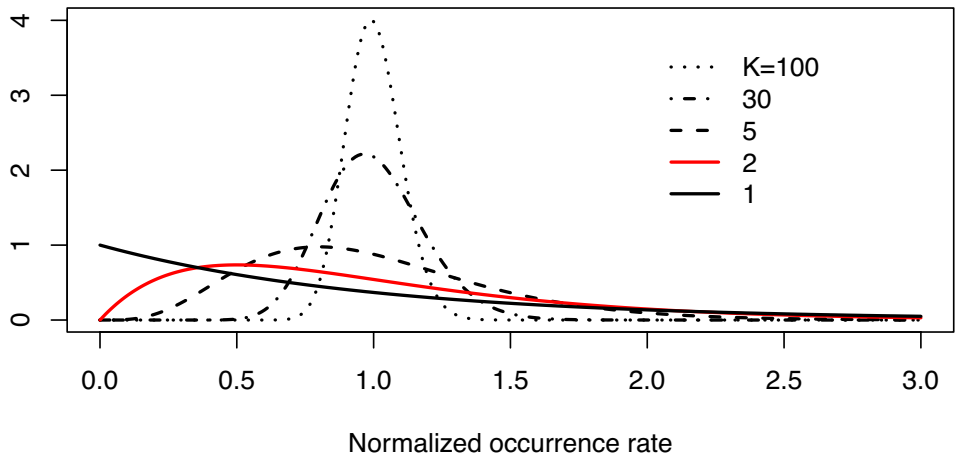


Figure 3 Degree of experience governing the concentration of the density of occurence rate

Thus, by combining Eq.(12) and (13), we obtain the same relation as Eq.(10).

In order to evaluate it in practice for our obtaining sample, we use the following form

$$\frac{1}{K} \;=\; \frac{\nabla'\lambda \;\mathcal{I}^{-1}\;\nabla\lambda}{\lambda^2} \tag{14}$$

where $\mathcal{I}$ is the observed information matrix. The inverse of $\mathcal{I}$ is used in behalf of the variance-covariance matrix $V(\boldsymbol{\theta})$ for the estimation errors of parameters. For the theoretical purpose, as a substitute for the observed information matrix, we employ the Fisher's expected information matrix, which is symmetrically expressed as

$$\mathcal{I} \;=\; NA^{-1}\left\{\begin{array}{ccc} p & \frac{\Gamma(2+\xi)-p}{\xi} & \frac{p}{\xi}-q \\ \frac{1-2\Gamma(2+\xi)+p}{\xi^2} & \frac{1}{\xi}\left(\frac{\Gamma(2+\xi)-p}{\xi}+q-r\right) & \\ & \frac{\pi^2}{6}+\frac{1}{\xi}\left(\frac{p}{\xi}-2q\right)+r^2 & \end{array}\right\}A^{-1} \tag{15}$$

in case of a GEV distribution (Prescott and Walden, 1980) applied to the annual maximum value distribution without any covariation, which is named the stationary model. For abbreviation, we use

$$p \;=\; (1+\xi)^2\,\Gamma(1+2\xi),\;\; q \;=\; \Gamma(2+\xi)\left\{1+\psi(1+\xi)+\frac{1}{\xi}\right\},\;\; r \;=\; 1+\psi(1)+\frac{1}{\xi} \tag{16}$$

and a diagonal matrix for adjusting the scale:

$$A \;=\; \begin{pmatrix} \sigma & 0 & 0 \\ & \sigma & 0 \\ & & \xi \end{pmatrix} \tag{17}$$

For the gradient of the occurence against the GEV parameters, we have

$$\frac{\nabla'\lambda}{\lambda} \;=\; \frac{1}{\lambda}\left(\frac{\partial}{\partial\mu},\;\frac{\partial}{\partial\sigma},\;\frac{\partial}{\partial\xi}\right)\lambda \;=\; \left(\lambda^{\xi},\;\frac{1-\lambda^{\xi}}{\xi},\;\log\frac{1}{\lambda}-\frac{1-\lambda^{\xi}}{\xi}\right)A^{-1} \tag{18}$$

Straightforwardly we can apply the manner above to the annual maximum value distribution with several covariates, the time and climate index that we are targeting on, which is named the non-stationary model. In this model, the covariates $\boldsymbol{x}'$ ($=\{x_1,\,x_2,\,\cdots x_m\}$) are linked to the GEV parameters in general form:

$$\mu \;=\; \boldsymbol{x}'\,\boldsymbol{\beta}_{\mu} \;=\; \mu_0+\beta_{\mu}^{(1)}x_1+\beta_{\mu}^{(2)}x_2+\cdots+\beta_{\mu}^{(m)}x_m \tag{19a}$$

$$\log\sigma \;=\; \boldsymbol{x}'\,\boldsymbol{\beta}_{\log\sigma} \;=\; \log\sigma_0+\beta_{\log\sigma}^{(1)}x_1+\beta_{\log\sigma}^{(2)}x_2+\cdots+\beta_{\log\sigma}^{(m)}x_m \tag{19b}$$

$$\xi \;=\; \boldsymbol{x}'\,\boldsymbol{\beta}_{\xi} \;=\; \xi_0+\beta_{\xi}^{(1)}x_1+\beta_{\xi}^{(2)}x_2+\cdots+\beta_{\xi}^{(m)}x_m \tag{19c}$$

The fisher information matrix becomes

$$\mathcal{I} \;=\; \begin{pmatrix} \mathcal{I}_0 & \mathbf{1}_N'\boldsymbol{X}_{-}\otimes\mathcal{I}_1 \\ \boldsymbol{X}_{-}'\mathbf{1}_N\otimes\mathcal{I}_1' & \boldsymbol{X}_{-}'\boldsymbol{X}_{-}\otimes\mathcal{I}_2 \end{pmatrix} \tag{20}$$

where $\otimes$ stands for the Kronecker product, and $\boldsymbol{X}_{-}$ is a matrix for sample covariates of m components over N years

$$\boldsymbol{X}_{-} \;=\; \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ \vdots & \vdots & \cdots & \vdots \\ x_{i,1} & x_{i,2} & \cdots & x_{i,m} \\ x_{N,1} & x_{N,2} & \cdots & x_{N,m} \end{pmatrix} \tag{21a}$$

For manipulation we use the matrices that consist of ones as

$$\mathbf{1}_N' \;=\; (\;\underbrace{1\;\cdots\;1}_{N}\;) \tag{21b}$$

and $\mathbf{1}_{(.)}$ is taken for extracting the subset $\mathcal{I}_1 = \mathcal{I}_0 \mathbf{1}_{(.)}$ and $\mathcal{I}_2 = \mathbf{1}'_{(.)} \mathcal{I}_0 \mathbf{1}_{(.)}$ from the Fisher's information matrix $\mathcal{I}_0$ as one of the followings:

$$\mathbf{1}_{(\mu)} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{1}_{(\log\sigma)} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{1}_{(\xi)} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad \mathbf{1}_{(\mu,\log\sigma,\xi)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{22a}$$

$$\mathbf{1}_{(\mu,\log\sigma)} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{1}_{(\mu,\xi)} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{1}_{(\log\sigma,\xi)} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{22b}$$

The degree of experience for non-stationary model with several covariates seem to be too much complicated to obtain any simple relation. Some algebraic formulae about the Kronecker product and the Schur complement helps us to obtain the decomposed form:

$$\frac{1}{K} = \frac{1}{K_0} + \frac{N}{M} \frac{MD^2(\boldsymbol{x})}{N-1} \tag{23}$$

where the degree of experience for non-stationary model K is separated from that for the stationary model K_0 and a modulus for the components of the GEV parameters linked to the covariates:

$$\frac{1}{K_0} = \frac{\nabla'\lambda \ \mathcal{I}_0^{-1} \ \nabla\lambda}{\lambda^2}, \qquad \frac{1}{M} = \frac{\nabla'_{(.)}\lambda \ \mathcal{I}_2^{-1} \ \nabla_{(.)}\lambda}{\lambda^2} \tag{24}$$

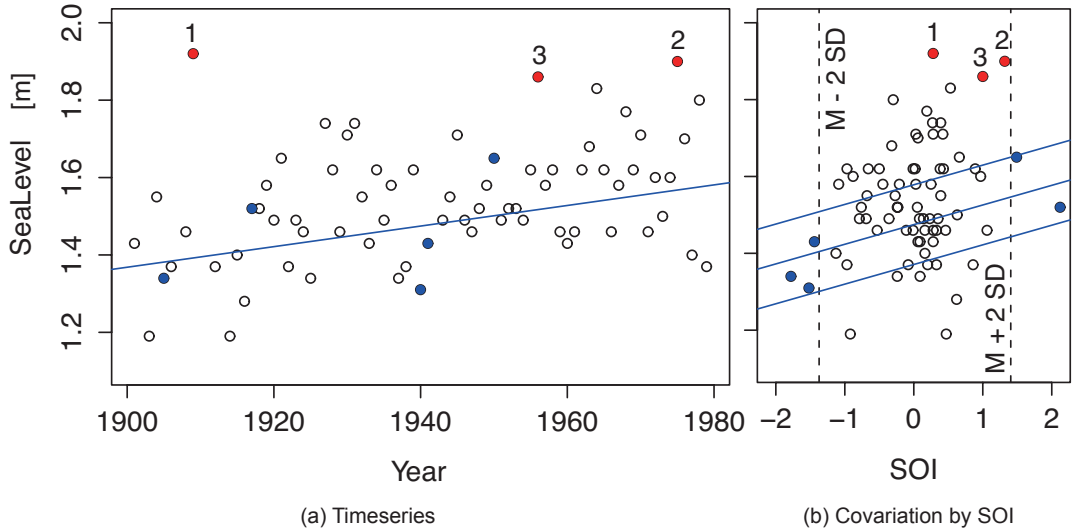
It is denoted that $\nabla_{(.)} = \mathbf{1}'_{(.)}\nabla$ for the subset of the information matrix, and in addition it should be noted carefully that $MD^2(\boldsymbol{x})$ means the Mahalanobis squared distance (see, for example, Weisberg, 1987) and it can be related to the leverage as

$$h_{ii} = \frac{1}{N} + \frac{MD^2(\boldsymbol{x}_i)}{N-1} \tag{25}$$

It is noted that the Mahalanobis squared distance can be evaluated for any other values of covariates than the observed ones in the advantage to the leverages which are originally difined as the diagonal components of the following matrix

$$\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'; \qquad \boldsymbol{X} = \{\mathbf{1}_N, \boldsymbol{X}_-\} \tag{26}$$

As seen in Eq.(23) and (25), the degree of experience for the non-stationary model plays the same role as the Cook's distance, by taking into consideration that the degree of experience for the stationary model shows the pure statistical variation for extremes which corresponds with the residuals in ordinary regression analysis. For the high leverage, the degree of experience for the non-



(a) Timeseries (b) Covariation by SOI

Figure 4 Annual maximum sea levels in Fremantle port

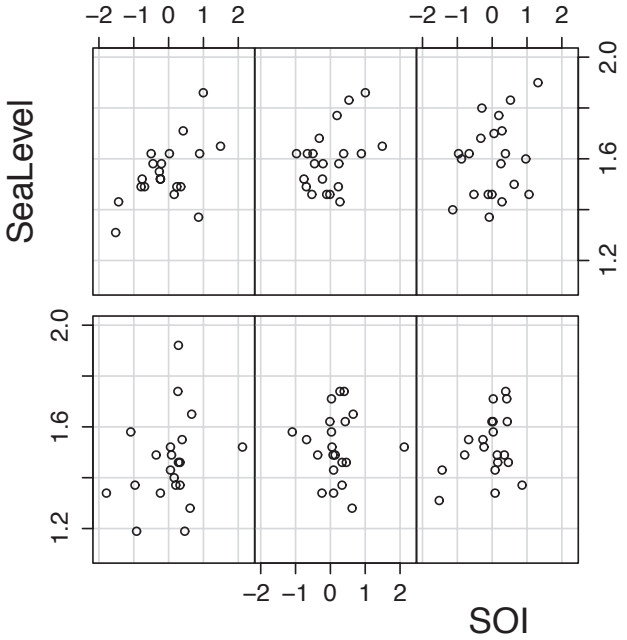
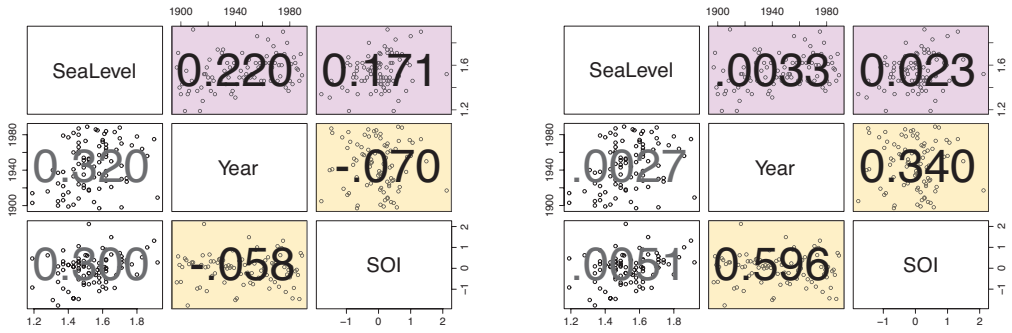


Figure 5 Splited scatter plots for sea levels with covariated SOI by splited time intervals

stationary model decreases but the direction of increment is opposite as the Cook's distance increases for high leverage. It should be remaked that the degree of experience can be decomposed into the two parts: just statistical deviation and the leverage due to the covariates' deviation. The former deviation is happened probably so it shouldn't be removable outlier, while the latter deviation is the fault of rare condition against the covariates and it should be removed as an influential outlier.

Illustrated demonstration by an example Fremantle sea levels

Flemantle port is located in west part of Australia, and it is used as an example covariated with the SOI in the text book by Coles (2001). Here we employ the data to demonstrate the diagram for detecting the influenced outlier. The timeseries of the sea levels over 79 years are shown in Fig. 4(a), where the red points are the largest three levels in the record and most of the blue points are mediocrity in the timeseries but they are outside of twice times as larger as the standard deviation against the SOI variations in Fig.4(b). There are drawn three tendency (regression) lines against SOI, which are different from the specified years. In the splited time intervals the scatted plots are



a) Values of correlation coefficients b) P-values for the significance

Fig. 6 Correlations among the extremes and the covariates
(the upper triangle by the product moment correlation, the lower triangle by the rank correlation)

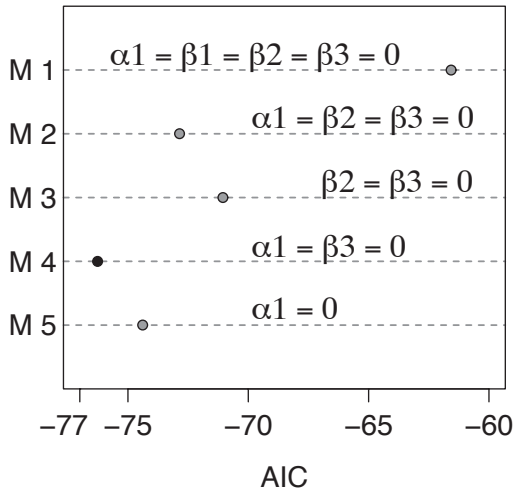


Figure 7 Model selection

drawn in Fig.5, which would tell us the correlation be suspicious, and the influenced outlier may pull the tendency lines against the covariate SOI. We clear the doubt by means of the outlier sensitivity diagram for extremes.

Fig. 6(a) shows the matrix of the paird scatter plots, those numbers indicates the values of the product moment correlation in the upper triangle and those of the Kendall's rank correlation in the lower triangle, and the Fremantle sea level has a temporal trend and covariated with SOI though their covariations are weak because the p-values are taken enough small to be significant as shown in the same positions of each triangle in Fig. 6(b). It is also found that both covariates has almost no correlation, which would make the problem so easier to approach in our thought.

As shown in Fig. 7, we examined five models by fixing the value of coefficients to zero in the following links to the covariates:

$$\mu = \mu_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \tag{27a}$$

$$\log \sigma = \log \sigma_0 + \alpha_1 x_1 \tag{27b}$$

$$\xi = \xi_0 \tag{27c}$$

According to the rule of thumb of AIC, we choose M4, whose value of coefficients are listed as

$$\alpha_1 = \beta_3 = 0; \quad \hat{\mu}_0 = 1.47, \ \hat{\beta}_1 = 0.1037, \ \hat{\beta}_2 = 0.0511, \ \hat{\sigma}_0 = 0.124, \ \hat{\xi}_0 = -0.15 \tag{28}$$

The degree of experience comes to the largest value 72, which corresponds to the actual sample size, at the relatively low level around 1.4 m, as seen in Fig. 8. The values of degree of experience are spreading even for the same sea level, because of taking the covariation due to both of SOI and time into consideration. Fig. 8 shows the degree of experience for the data in the two manner. One is maked by the open circle, which is the degree of experience obediently evaluated by Eq.(14) with the observed information matrix given by the sample. Another is marked by gray color, which is a kind of approximation evaluated by Eq.(23). Those by Eq.(14) agree well with those by Eq.(23). Thus, those decomposed two terms in Eq.(23) can be regarded to be derived from the obedient evaluation by Eq.(14).

Hence in Fig. 9, we have the contingent discrepancy K_0 and the (inversed) leverage against the sea level, respectively. Several ones of the blue points, which are out of twice times the standard deviation of SOI as mentioned in Fig. 4.(b), take high leverage value, the inverse of which and are less than 12.0 in Fig. 9 as well as one of the red points. The critical value for the inverse of leverage is proposed as $N/2p$ (see Gross, 2003, for example), which gets $72/2/3 = 12.0$ in this case. These are candidates of influencial outlier.

By putting toghether the contingent discrepancy K_0 and the (inversed) leverage against the sea level, we have a diagram of the outlier sensivity Fig. 10, where the contour lines of the theoretical (approximated) values in the degree of experience in the horizontal axis of the reciproal number of

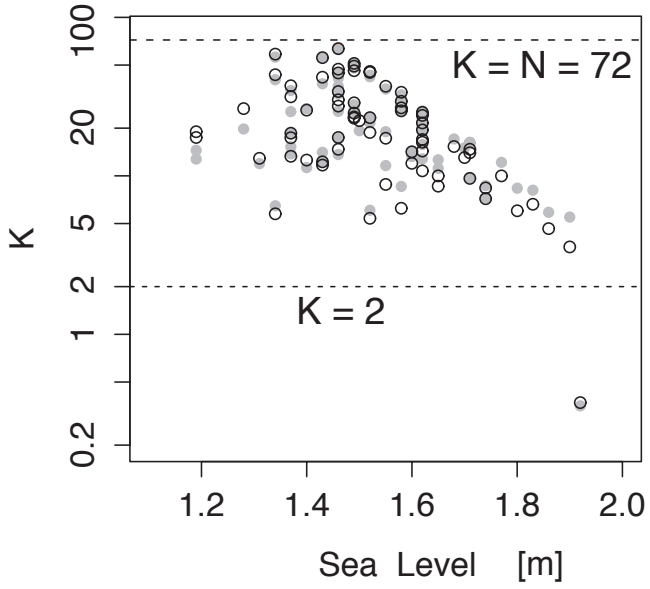


Figure 8 Two manners for evaluating the degree of experience

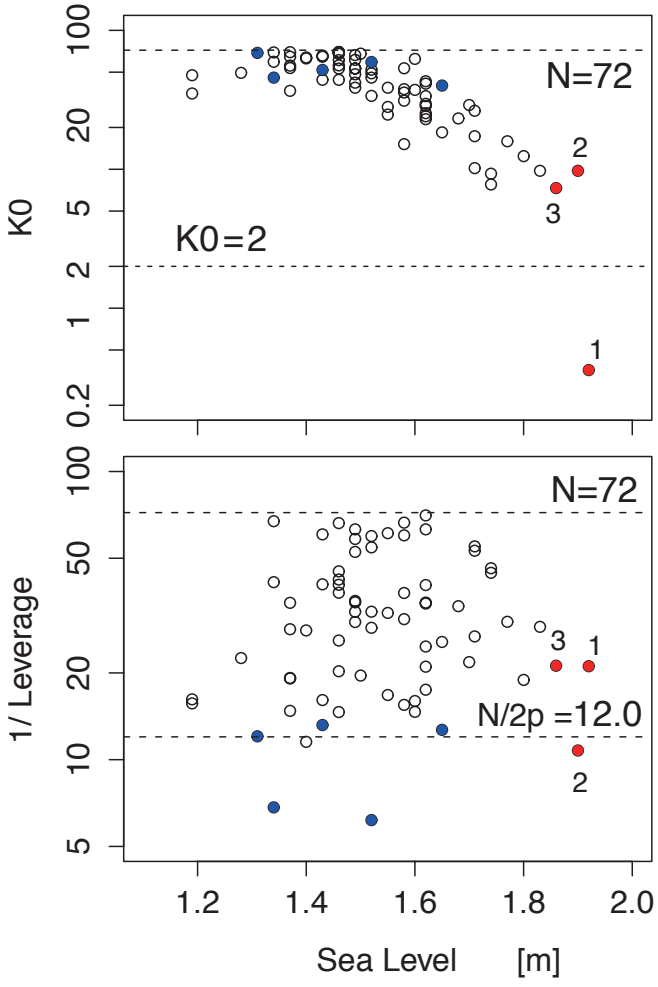


Figure 9 Two components in degree of experience

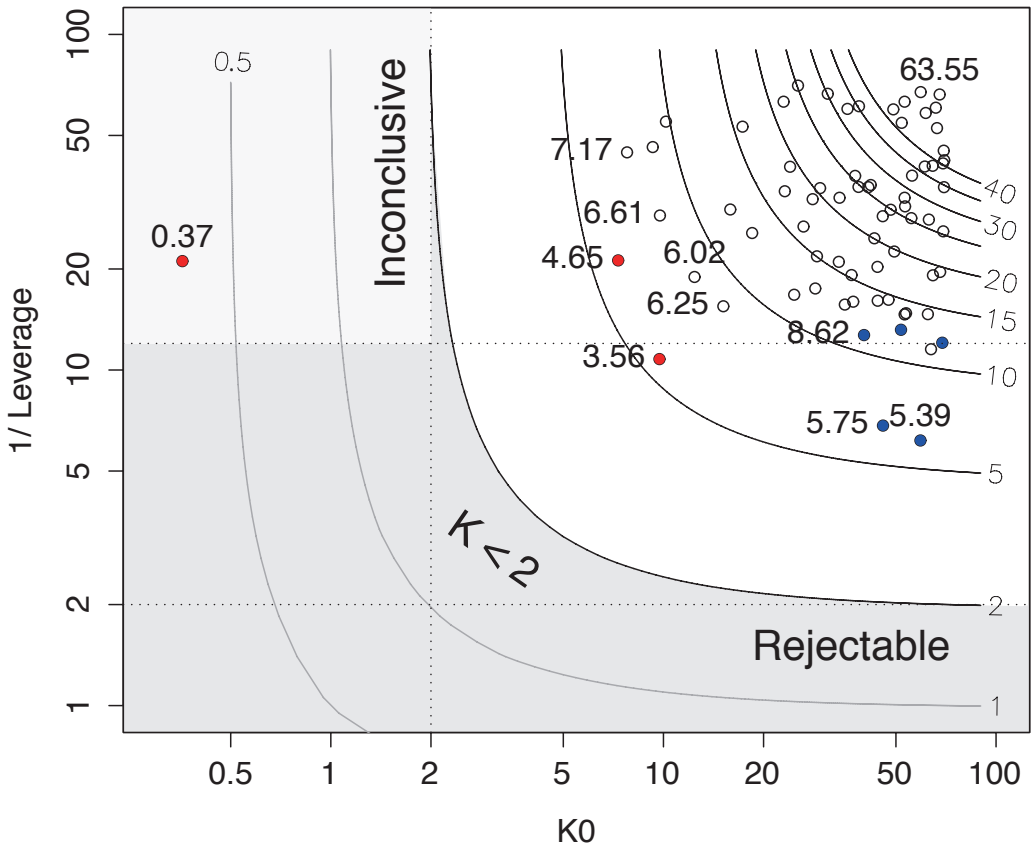


Figure 10 Outlier sensitivity diagram for extreme value analysis

leverages and the vertical axis of the degree of experience by no covariance model K_0.

In principle, the outliers, whose value of degree of freedom less than 2, are inconclusive. It means that those are acceptable as is, and those are included for the extreme analysis, but the results for those should not be concerned (e.g. the extremely largest value estimated for the return period of the record maximum should not surprised, because the result for those outliers are inconclusive. It is not correct nor wrong.) However, we have an exception: the outliers of high leverage (the reciprocal number is less than 2) should be removal. It is because the conditions of the covariate SOI and time are restricted. Fortunately, we have no outlier to remove. The red point, whose value less than 2 but low leverage, is just inconclusive. Five blue points are found to be enough low leverage, though those SOI are deviated from the others as seen in Fig. 4(b) and 9. The critical valus of degree of experience $K = 2.0$ is adopted also here after the proverbial reason, shown in Kitano et al. (2009), what happend twice will happen three times.

Conclusions

A diagram drawn by the contours for the degree of experience with the two axes of the inversed leverage and the contingent discrepancy is proposed in this study. It is based on the mathematical derivation of the decomposed form by the Fisher's information matrix. It is possible to detect the influenced outlier that the degree of experience is smaller because of high leverage, while we cannot reject the candidate outlier whose degree of experience takes a small value though the leverage is not so high. We should keep in suspence to judge the rejection because the outlier is just deviated occassionally. It will become more difficult to detect the influenced outlier in the dataset of more higher dimension of covariates for the research on the climate change. In such cases we hope that the detection method by the degree of experience in the diagram proposed here will be served usefully.

ACKNOWLEDGMENTS

This work was partiallly supported by Grant-in-Aid for Scientific Research (C) 21560538.

REFERENCES

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, Vol. 175, Springer, 394p.

Kitano, T., W. Kioka and R. Takahashi (2008): Degree of experience in statistical analysis for extreme wave heights, Ann. Jour. of Coastal Eng., JSCE, Vol. 55, pp.141-145. (in Japanese)

Kitano, T., W. Kioka and R. Takahashi (2009): Degree of experience for extreme wave statistics, Proc. of Coastal Dynamics 2009, World Scientific, Paper #209 (in CDROM).

Kitano, T., W. Kioka and R. Takahashi (2010): Trend model of sea extremes, Proceedings of 32nd Conference on Coastal Engineering, Shanghai, China, ASCE, paper #1169, 13p.

Kitano, T., W. Kioka and R. Takahashi (2011): Diffractive uncertainty toward the future estimation of return wave height, Coastal Structures 2011, Yokohama, ASCE, paper #C7-098, 11p.

Prescott, P. and A. T. Walden (1980): Maximum likelihood estimation of the parameters of the generalized extreme-value distribution, Biometrika, Vol.67, pp.723-724.

Weisberg, S. (1985): Applied Linear Regression, Wiley, 324p.