Extreme Event Modelling

Liwei Wu
Supervisor: Dr. Xiang Zhou

April 28, 2014

Table of Contents

Introduction

- Extreme Event: event occurring very rarely, such as the 100-year floor.
- We need the Extreme Event Modelling techniques to know what the extreme levels can be within a certain period of time.
- For example, the past 10 years of data $\rightarrow$ predict how large the 100-year flood can be

Two Methods available

1. Block Maxima Method
2. Threshold Method

Application to real data

1. Dow Jones Index data
2. Hong Kong climate data

- The models introduced later focuses on the statistical behavior of

$$M_n = \max\{X_1, X_2, ..., X_n\},$$

  where $X_1, X_2, ..., X_n$, is a sequence of independent random variables having a common distribution function F.
- In Applications, the X_i usually represents values of a process measured on a regular time-scale – perhaps hourly measurements of sea-level, or daily mean temperatures
- M_n represents the maximum of the process over n time units of observation.
- If n is the number of observations in a year, then M_n corresponds to the annual maximum.

Table of Contents

Table of Contents

Extremal Types Theorem

Theorem

If there exist sequences of constants $\{a_n\}$ *and* $\{b_n\}$ *such that*

$$P\left(\frac{M_n - b_n}{a_n} \leq z\right) \to G(z) \text{ as } n \to \infty$$

where G is a non-degenerate distribution function, then G belongs to one of the following families:

I

$$G(z) = \exp\{-\exp\left[-\left(\frac{z-b}{a}\right)\right]\}$$

II

$$G(z) = \begin{cases} 0 & \text{if } z \leq b; \\ \exp\{-\left(\frac{z-b}{a}\right)^{-\alpha}\} & \text{if } z > b. \end{cases}$$

III

$$G(z) = \begin{cases} \exp\{-\left(\frac{z-b}{a}\right)^{-\alpha}\} & \text{if } z \leq b; \\ 0 & \text{if } z > b. \end{cases}$$

for parameters $a > 0$, *b and, in the case of families II and III,* $\alpha > 0$.

- Collectively, these three classes of distribution are termed the extreme value distributions, with types I, II and III widely known as the Gumbel, Fréchet and Weibull families respectively.
- an extreme value analog of the central limit theorem.

The Generalized Extreme Value Distribution

Theorem

If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$P\left(\frac{M_n - b_n}{a_n} \leq z\right) \rightarrow G(z) \text{ as } n \rightarrow \infty \tag{1}$$

for a non-degenerate distribution function G, then G is a member of the GEV family

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}$$

defined on the set $\{z : 1 + \xi(z-\mu)/\sigma > 0\}$, where $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$.

Modelling extremes using Block Maxima

A series of independent observations X_1, X_2...are blocked into sequences of observations of length n, for some large value of n, generating a series of block maxima, $M_{n,1}, ..., M_{n,m}$, say, to which the GEV distribution can be fitted.

- Return Period
- Return Level

z_p is the return level associated with the return period $1/p$, which means the level z_p is expected to be exceeded on average once every $1/p$ years. Define $y_p = -log(1-p)$, then we have

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - y_p^{-\xi}\right], & \text{for } \xi \neq 0; \\ \mu - \sigma log\, y_p, & \text{for } \xi = 0. \end{cases}$$

Return Level Plot

The **return level plot** is the graph in which if Z_p is plotted against y_p on a logarithmic scale.

- If $\xi = 0$, the plot is linear.
- If $\xi < 0$, the plot is convex with asymptotic limit as $p \to 0$ at $\mu - \sigma/\xi$.
- If $\xi > 0$, the plot is concave and has no finite bound.

Block Maxima Method

1. Blocking the data into blocks of equal length. (Trade-off between bias and variance)
2. Fitting the GEV to the set of block maxima $Z_1, ..., Z_m$.
3. Obtaining the log-likelihood for the GEV parameters. When $\xi \neq 0$,

$$l(\mu, \sigma, \xi) = -m\log(\sigma) - (1+1/\xi)\sum_{i=1}^{m}\log\left[1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right] - \sum_{i=1}^{m}\left[1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right]^{(-1/\xi)}, \quad (2)$$

provided that

$$1+\xi\left(\frac{z_i-\mu}{\sigma}\right) > 0, \forall i = 1, .., m \quad (3)$$

4. Maximization of the equation with respect to the parameter vector (μ, σ, ξ) leads to the maximum likelihood estimate of the parameters.
5. Obtaining the maximum likelihood estimate of z_p for the 1/p return level.
6. Using the delta method to obtain the variance of the maximum likelihood estimate

$$Var(\hat{z}_p) \approx \nabla z_p^T V \nabla z_p$$

$$\begin{aligned}\nabla z_p^T &= \left[\begin{array}{ccc} \frac{\partial z_p}{\partial \mu} & \frac{\partial z_p}{\partial \sigma} & \frac{\partial z_p}{\partial \xi} \end{array}\right] \\ &= \left[\begin{array}{ccc} 1 & -\xi^{-1}\left(1-y_p^{-\xi}\right) & \sigma\xi^{-2}\left(1-y_p^{-\xi}\right) - \sigma\xi^{-1}y_p^{-\xi}\log y_p \end{array}\right]\end{aligned}$$

evaluated at $\left(\hat{\mu}, \hat{\sigma}, \hat{\xi}\right)$.

Modelling Checking

- Probability Plot: a comparison of the empirical and fitted distribution functions

$$\left\{\left(\tilde{G}\left(z_{(i)}\right), \hat{G}\left(z_{(i)}\right)\right), i = 1, ..., m\right\}$$

- Quantile Plot

$$\left\{\left(\hat{G}^{-1}\left(i/(m+1)\right), z_i\right), i = 1, ..., m\right\}$$

- Return Level Plot summarises the fitted model and consists of the locus of points

$$\left\{\left(\log\ y_p, \hat{z}_p\right), 0 < p < 1\right\}$$

  Confidence intervals can be added to the plot to increase its informativeness. Empirical estimates of the return level function can also be added, enabling the return level plot to be used as a model diagnostic.
- Density Plot: a comparison of the probability density function of a fitted model with a histogram of the data

We mainly check whether the probability plot and the quantile plot are approximately linear.

GEV distribution for Minima

Definition

$$\tilde{M}_n = \min\{X_1, ..., X_n\}$$

and

$$\tilde{\mu} = -\mu$$

Theorem

If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$P\left(\frac{\tilde{M}_n - b_n}{a_n} \leq z\right) \rightarrow \tilde{G}(z) \text{ as } n \rightarrow \infty \tag{4}$$

for a non-degenerate distribution function $\tilde{G}$, then $\tilde{G}$ is a member of the GEV family of distributions for minima:

$$\tilde{G}(z) = 1 - \exp\left\{-\left[1 - \xi\left(\frac{z-\tilde{\mu}}{\sigma}\right)\right]^{-1/\xi}\right\}$$

defined on the set $\{z : 1 + \xi(z - \tilde{\mu})/\sigma > 0\}$, where $-\infty < \tilde{\mu} < \infty$, $\tilde{\sigma} > 0$ and $-\infty < \xi < \infty$.

Table of Contents

Motivation

Modelling only block maxima is a wasteful approach to extreme value analysis if other data on extremes are available. Therefore, if an entire time series of, say, hourly or daily observations is available, then we can make better use of the data by avoiding altogether the procedure of blocking. Let $X_1, X_2, ...$ be a sequence of independent random variables with common distribution function F, and let

$$M_n = \max\{X_1, ..., X_n\}$$

It is natural to regard those of the X_i exceeding some high threshold u as extreme events.

The Generalized Pareto Distribution

Theorem

Denote an arbitrary term in the X_i sequence by X, and suppose that F satisfies Theorem 2.2, so that for large n,

$$P\left(M_n \leq z\right) \approx G(z),$$

where

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}$$

for some $\mu, \sigma > 0, \xi$. Then, for large enough u, the distribution function of $(X-u)$, conditional on $X > u$, is approximately

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi} \tag{5}$$

defined on $\{y : y > 0$ and $(1 + \xi y/\tilde{\sigma}) > 0\}$, where

$$\tilde{\sigma} = \sigma + \xi(u-\mu)$$

Threshold Selection

The issue of threshold choice is analogous to the choice of block size in the block maxima approach

- too low a threshold is likely to violate the asymptotic basis of the model, leading to bias
- too high a threshold will generate few excesses with which the model can be estimated, leading to high variance.

Two methods for Threshold Selection

- Mean residual life plot
- Estimate the model at a range of thresholds.

Model Checking

After obtaining the proper threshold of the fitted GPD, we need to assess the quality of the fitted generalized Pareto model. It can be done using probability plots, quantile plots, return level plots, and density plots.

Note that the return level plot is given by $\{(m, \hat{x}_m)\}$ for large values of m, where

$$\hat{x}_m = u + \frac{\hat{\sigma}}{\hat{\xi}}\left[\left(m\hat{\zeta}_u\right)^{\hat{\xi}} - 1\right]$$

and

$$\hat{\zeta}_u = Pr\{X > u\}$$

Table of Contents

Table of Contents

Simulation of data from normal distribution

- Consider the data generated from R $x_{1,1}, x_{1,2}, ..., x_{1,n}, x_{2,1}, x_{2,2}, ..., x_{2,n}, x_{3,1}, ..., x_{m,1}, ..., x_{m,n}$ from standard normal distribution: $\mathcal{N}(0,1)$.
- Dividing the data into m blocks, we have n data points in each block.
- We now simplify notation by denoting the block maxima $Z_1, Z_2, ..., Z_m$. These are assumed to be independent variables from a GEV distribution whose parameters are to be estimated given that n is large enough.
- In this specific case, we take $m = 200$, $n = 500$.
- Fitting against GEV distribution vs. Gumbel distribution.

GEV distribution

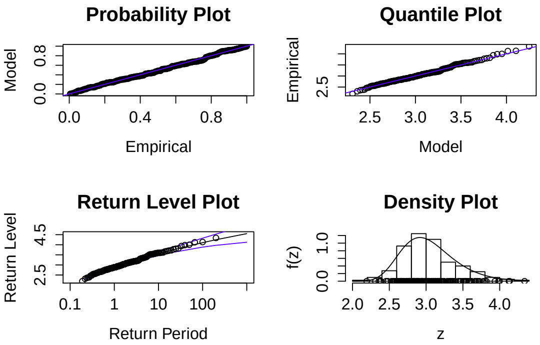

Figure : GEV fit

Using packages in R, we have the estimates as follows: $\mu = 2.886553$, with a 95% confidence interval $[2.837136, 2.93597]$, $\sigma = 0.3216$, with a 95% confidence interval $[0.286984, 0.3562161]$, $\xi = -0.08799757$, with a 95% confidence interval $[-0.1778897, 0.001894544]$.

Gumbel distribution

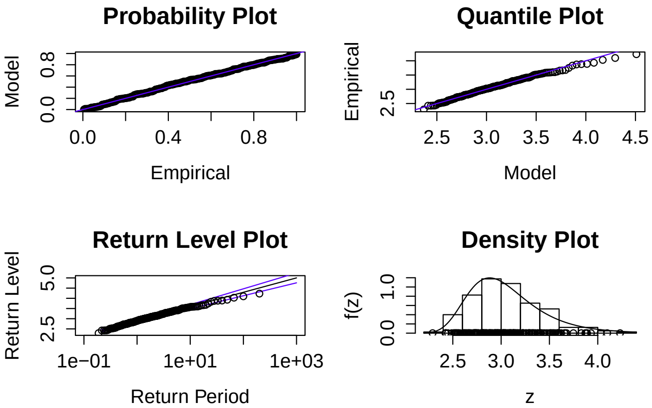

Figure : Gumbel fit

Conclusion

- Notice that the confidence interval of ξ contains 0, which means the Gumbel Distribution could be a more accurate model in the entire GEV family.
- Confidence intervals for return levels obtained by fitting against Gumbel Distribution are narrower than those obtained by fitting against a member of general GEV distribution.
- For the specific example above, estimates and confidence intervals for returns levels can be obtained. For example, when $p = 1/10$, the estimate for the 10-year return level $\hat{z}_{0.1}$is 3.536803, with a 95% confidence interval $[3.453434, 3.641219]$, and similarly when $p = 1/100$, the estimate for the 100-year return level $\hat{z}_{0.01}$is 4.136239, with a 95% confidence interval $[3.955557, 4.447778]$.

Table of Contents

Application in stock market

- Sometimes people are interested in the extreme event in stock markets, such as how large the daily loss percentage would be within 100 years. Such problems can be modelled and solved using **Threshold Method** from the Extreme Event Modelling.
- daily Dow Jones Index data from 1896-05-26 to 2013-08-02, totalling 31960 data points
- After inputting the csv file containing the data into R, we can use the existing R package "ismev" and "extRemes" to choose the proper threshold, fit the data into the Generalized Pareto Distribution (GPD), and obtain the 100-year return level of daily loss percentage in stock market with the corresponding 95% confidence interval.

Data points Plot

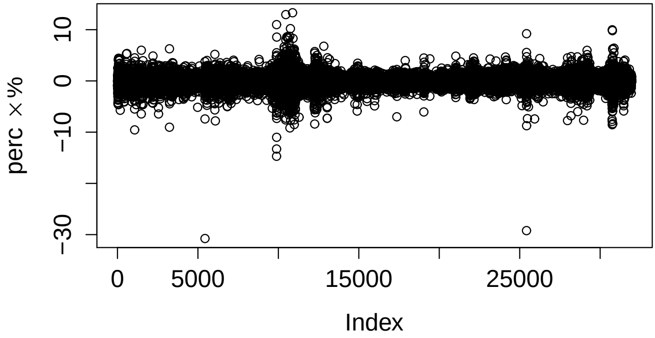

Figure : data points of daily percentage change of Dow Jones Index

Selecting the proper threshold: First Method

- The first method is to select a proper threshold such that the mean residue life plot should be approximately linear above the selected proper threshold u_0.
- In the mean residue life plot below, the information above the daily loss percentage $u = 10$ is not very accurate due to very few points (actually only 5) with daily loss greater than 10%. Therefore, we should ignore that part of the plot and conclude that the proper threshold u_0 should satisfy $u_0 \geq 3$, since it is easy to see that the plot is approximately linear above $u = 3$.

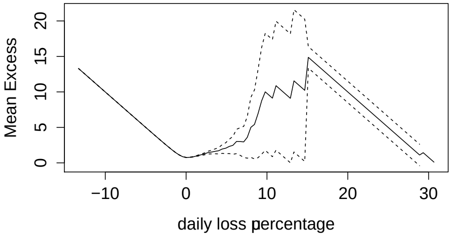

Figure : mean residue life plot

Selecting the proper threshold: Second Method

- To further explore what the proper threshold should be, the second method is used: look for the stability of parameters σ^* and ξ while varying the threshold of the fitted GPD.
- From the plot on the next slide, we can see that the estimated parameters are more or less stable when $u \geq 5$. Therefore, the selected threshold of $u = 5$ appears reasonable.

Selecting the proper threshold: Second Method

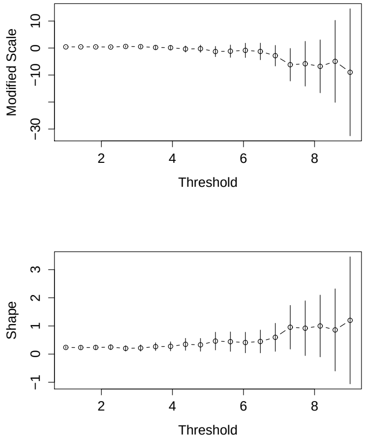

Figure : parameter estimates against threshold

Obtaining the 100-year return level of daily loss percentage

1. Fit the data points into the GPD with the selected threshold $u = 5$
2. Obtain the maximum likelihood estimates in this case:

$$\left(\hat{\sigma}, \hat{\xi}\right) = (1.2694261, 0.3737744)$$

   with a corresponding maximised log-likelihood of -143.4933. Standard errors for $\hat{\sigma}$ and $\hat{\xi}$ are 0.2096583 and 0.1327465 respectively.
3. Obtain the 100-year return level $\hat{x}_m$: $\hat{x}_m = 20.710722$ and $Var\left(\hat{x}_m\right) = 27.80601$, leading to a 95% confidence interval for x_m of $[10.375366, 31.04608]$.
4. Draw the diagnostic plots for model checking (shown on next slide)

Model Checking

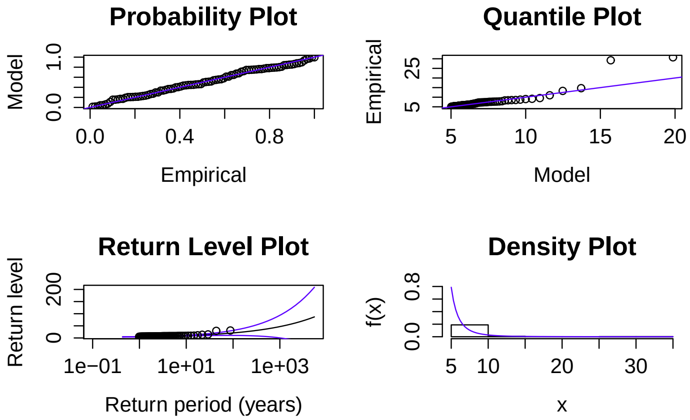

Figure : Diagnostic plots

Return level plot

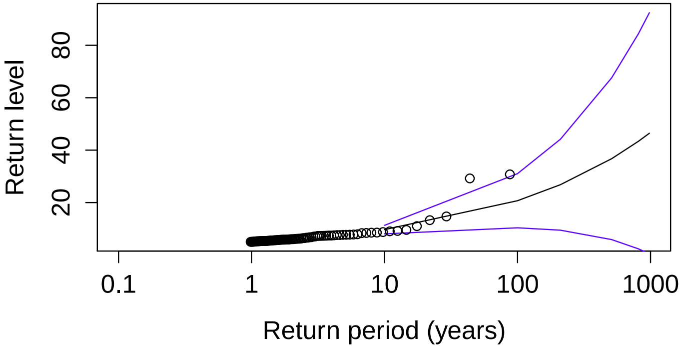

Figure : return level plot

Conclusion

- we are 95% confident to say that the maximum daily loss percentage within a hundred years will fall between 10.375366 and 31.04608.
- perhaps some precautions can be done to get prepared for the possible stock market crash given that we already know how large the maximum daily loss percentage can be within a hundred years.
- In particular, the same statistical method used here for Dow Jones Index can be applied in other stock market around the world, such as Hong Kong, China, Japan and so on. The comparisons of the 100-year return level among these different stock markets would be interesting. In addition, the analysis of data at financial turmoil period could reveal extreme patterns of financial market under huge uncertainty.

Table of Contents

Application in Hong Kong climate data

- It is a long tradition to use the extreme value analysis to study climate and meteorology problem. Here we study the daily maximum temperature in Hong Kong.
- the Hong Kong climate data is publicly available from the Hong Kong Observatory website: `http://www.weather.gov.hk/cis/data_e.htm`.
- we will only focus on the maximum air temperature, because our goal is to explore how high the maximum air temperature can be within a hundred years and within a hundred and fifty years respectively.

Threshold Method

- we have to take into account the non-stationarity of the data points, since obviously the daily maximum air temperature will depend on the month of the year and tend to cluster together
- To overcome the difficulty, we will only select the data points from June, July and August, since by observation only the data points in these three months will likely become the maximum air temperature throughout the year. We draw the plot of the data points.
- We obtained the daily maximum air temperature data in June, July and August from 1997 till 2013, totalling 1564 data points. We can use the existing R package "ismev" and "extRemes" to choose the proper threshold, fit the data into the Generalized Pareto Distribution (GPD), and obtain the 100-year and 150-year return level of daily maximum air temperature in Hong Kong with the corresponding 95% confidence interval.

Data points Plot

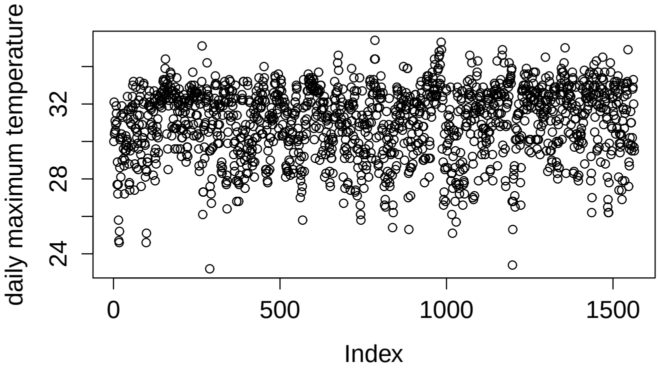

Figure : data points of daily maximum temperature of Hong Kong climate data

Selecting the proper threshold: First Method

- The first method is to select a proper threshold such that the mean residue life plot should be approximately linear above the selected proper threshold u_0.
- In the mean residue life plot below, it is approximately linear above $u = 32$. Therefore, we can conclude that the proper threshold u_0 should satisfy $u_0 \geq 32$.

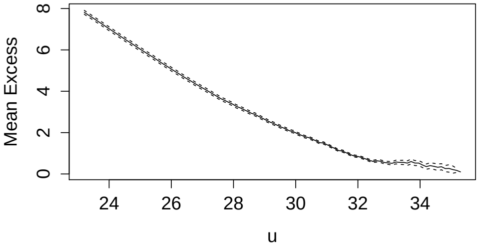

Figure : mean residue life plot

Selecting the proper threshold: Second Method

- To further explore what the proper threshold should be, the second method is used: look for the stability of parameters σ^* and ξ while varying the threshold of the fitted GPD.
- From the plot on the next slide, we can see that the estimated parameters are more or less stable when $u \geq 32$. Therefore, the selected threshold of $u = 32$ appears reasonable.

Selecting the proper threshold: Second Method

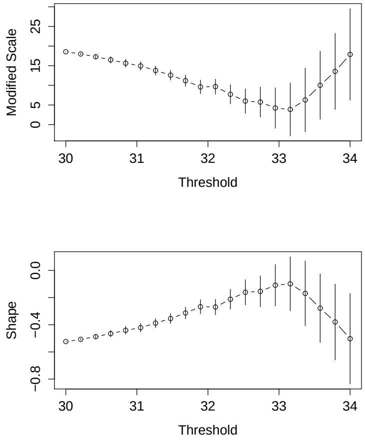

Figure : parameter estimates against threshold

Obtaining the 100-year return level of daily loss percentage

1. Fit the data points into the GPD with the selected threshold $u = 32$
2. Obtain the maximum likelihood estimates in this case:

$$\left(\hat{\sigma}, \hat{\xi}\right) = (1.0722162, -0.2809016)$$

   with a corresponding maximised log-likelihood of -449.7361. Standard errors for $\hat{\sigma}$ and $\hat{\xi}$ are 0.05182593 and 0.02684719 respectively.
3. Obtain the 100-year return level $\hat{x}_m$: $\hat{x}_m = 35.55201$ and $Var\left(\hat{x}_m\right) = 0.02606413$, leading to a 95% confidence interval for x_m of $[35.23559, 35.86844]$.
4. Obtain the 150-year return level $\hat{x}_m$: $\hat{x}_m = 35.58055$, and a 95% confidence interval for x_m of $[35.25327, 35.90782]$.
5. Draw the diagnostic plots for model checking (shown on next slide)

Model Checking

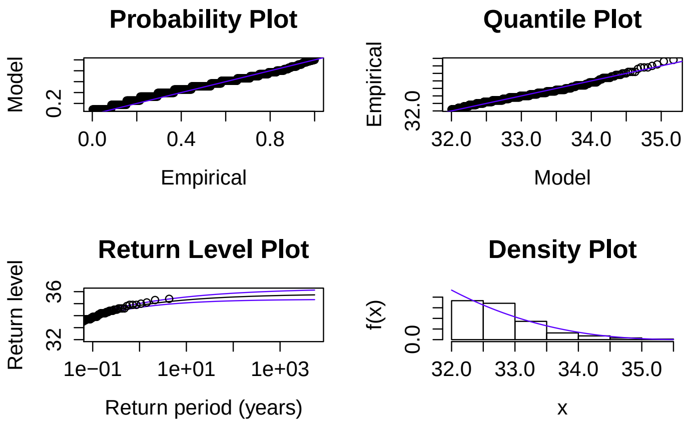

Figure : Diagnostic plots

Return level plot

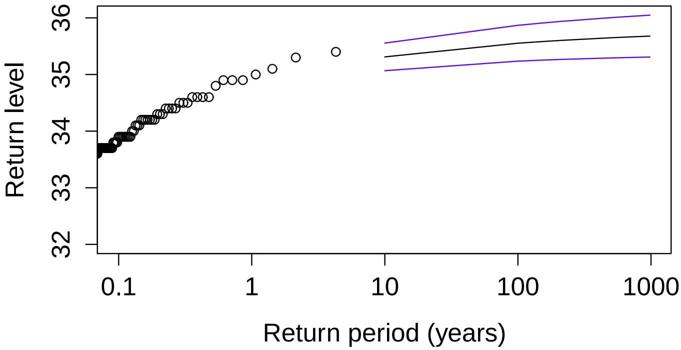

Figure : return level plot

Conclusion

- we are 95% confident to say that the maximum daily loss percentage within a hundred years will fall between 35.23559 and 35.86844.
- we are 95% confident to say that the maximum air temperature within a hundred and fifty years will fall between 35.25327 and 35.90782.
- Although the range is pretty wide, we can still conclude that the maximum air temperature will most likely never reach 36 degrees in Hong Kong even within 150 years. Therefore, we know what we are up to and take some precautionary measures in the event of extremely high air temperature.

Table of Contents

Summary

- we discussed the basic theory and methods of Extreme Value Modelling. Both the Generalized Extreme Value (GEV) distribution and the Generalized Pareto Distribution (GPD) are introduced.
- two methods to model the Extremes, i.e. the Block Maxima method and the Threshold method.
- we applied these methods to fit the data to obtain the corresponding return level.
  - we applied the Block Maxima Method to the simulated normal data.
  - we applied the Threshold Method to the Dow Jones Index data and the Hong Kong climate data and reached informative conclusion based on the return levels we obtained.

Outlook

- In the Hong Kong climate data case, we assume that the data are obtained from one station of the same location. However, in real world application, it is possible that the data are obtained from a number of stations of different locations.
- If so, we can use the Bayesian hierarchical model for spatial extremes to produce a map characterising extreme behaviour across a geographic region.
- There are three layers in both of our hierarchal models: Data Layer, Process Layer, & Priors
- $$p(\theta|Z(x)) \propto p_1(Z(x)|\theta_1)p_2(\theta_1|\theta_2)p_3(\theta_2)$$
- Based on the equation above, we can obtain the posterior distributions of $\sigma(x)$, $\xi(x)$, and $\zeta(x)$ by using MCMC algorithms. Then, the return level posterior distribution as well as the return level maps can be produced accordingly.