Intro to Extreme Value Analysis

Overview

In order to predict rare events, of particular interest in current times of accelerating climate change are the risks of natural disasters. These events are particularly important in the insurance field as they are difficult to predict but can have catastrophic consequences. A recent example could be Hurricane Milton which made landfall in Florida and one that everyone will remember is Hurricane Katrina in which the levees broke and caused catastrophic flooding with little warning causing billions in damages and the bankruptcy of multiple companies. Who can forget the vivid images of the city of New Orleans underwater and the people stranded on their rooftops waiting for rescue. Extreme value analysis is important to actuaries and is used to predict the probability of rare events where we have very few observations.

```
#load relevant libaries, use install.packages() if you do not have them installed
library(extRemes)
library(evd)
library(httpgd)
library(ggplot2)
library(ggthemes)
library(support.CEs)
library(evir)
library(dgof)
library(ROOPSD)
```

Block-Maxima Method

What are Block Maxima:

Lets define a series of random variables

$$X_1, X_2, X_3, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F$$

which are independent and identically distributed (i.i.d) with a common distribution function $F(x)$.

We want to look at the tails of this distribution so lets create a new series

$$M_n = max(X_1, X_2, X_3, \ldots, X_n)$$

which allows us to take the maximum per block of observations, i.e. for N observations ve would have N/m=k for k number of blocks of size m.

To define the distibution function of M_m we can say

$$P(M_n \leq z) = \mathbb{P}(X_1 \leq z, X_2 \leq z, \ldots, X_n \leq z)$$

and as all the X_i are independant we can represent this as

$$P(M_n \leq z) = \mathbb{P}(X_1 \leq z)\mathbb{P}(X_2 \leq z) \ldots \mathbb{P}(X_n \leq z)$$

since the cdf for each variable X_i is for each i we can use $F(z) = P(X_i \leq z)$ and we can write the above as

$$P(M_n \leq z) = F(z)^n$$

The Fisher - Tippet - Gnedenko Theorem suggests that there are 2 series of constants $a_n > 0$ and $b_n \in \mathbb{R}$ such that as $n \to \infty$. Here we see that this is analagous to the central limit theorem giving a limiting distribution for extremes.

Considering this we have

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq z\right) \to G(z)$$

where $G(z)$ is a Generalised Extreme Value (GEV) distribution function with

$$G(z) \propto \exp\left[-(1+\xi z)^{-\frac{1}{\xi}}\right]$$

It can be shown that G(z) will belong to one of 3 families of distributions depending on the value of the shape parameter ξ.:

Gumbel:

$$G(z) = \exp\left\{-\exp\left(-\frac{z-b}{a}\right)\right\}, \quad z \in \mathbb{R}$$

Weibull:

$$G(z) = \begin{cases} \exp\left\{-\exp\left(-\frac{z-b}{a}\right)\right\} & \text{if } z < b \\ 1 & \text{if } z \geq b \end{cases}$$

Frechet:

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

These distributions are useful in modelling the behaviour of the tails of a series of data with.

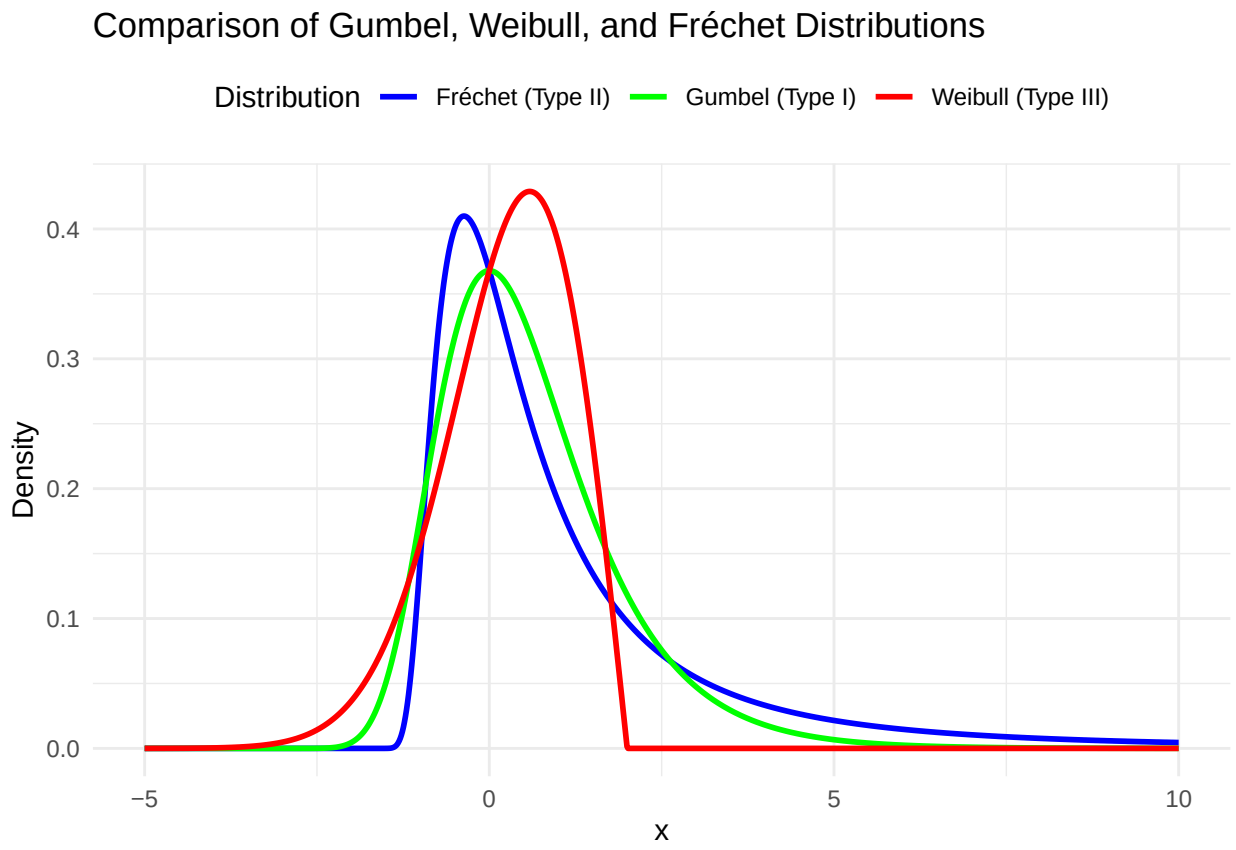
- **Gumbel (Type I):** $\xi = 0$, representing light-tailed distributions
- **Weibull (Type III):** $\xi < 0$, representing bounded distributions where extreme values have an upper limit
- **Fréchet (Type II):** $\xi > 0$, representing heavy-tailed distributions, which have a higher probability of extreme values

The parameter a_n, b_n, ξ are typically estimated using Maximum Likelihood Estimation (MLE). This can be done using the fgev function in R.

Below is a plot to visualise the differences between the Gumbel, Weibull, and Fréchet distributions.

```
x <- seq(-5, 10, length.out = 1000)

gumbel_vals <- devd(x, loc = 0, scale = 1, shape = 0, type = "GEV")
weibull_vals <- devd(x, loc = 0, scale = 1, shape = -0.5, type = "GEV")
frechet_vals <- devd(x, loc = 0, scale = 1, shape = 0.5, type = "GEV")


data <- data.frame(
```

```
  x = rep(x, 3),
  Density = c(gumbel_vals, weibull_vals, frechet_vals),
  Distribution = rep(c("Gumbel (Type I)", "Weibull (Type III)", "Fréchet (Type II)"), each = length(x))
)

ggplot(data, aes(x = x, y = Density, color = Distribution)) +
  geom_line(size = 1) +
  labs(title = "Comparison of Gumbel, Weibull, and Fréchet Distributions",
       x = "x", y = "Density") +
  theme_minimal() +
  theme(legend.position = "top") +
  scale_color_manual(values = c("blue", "green", "red"))
```

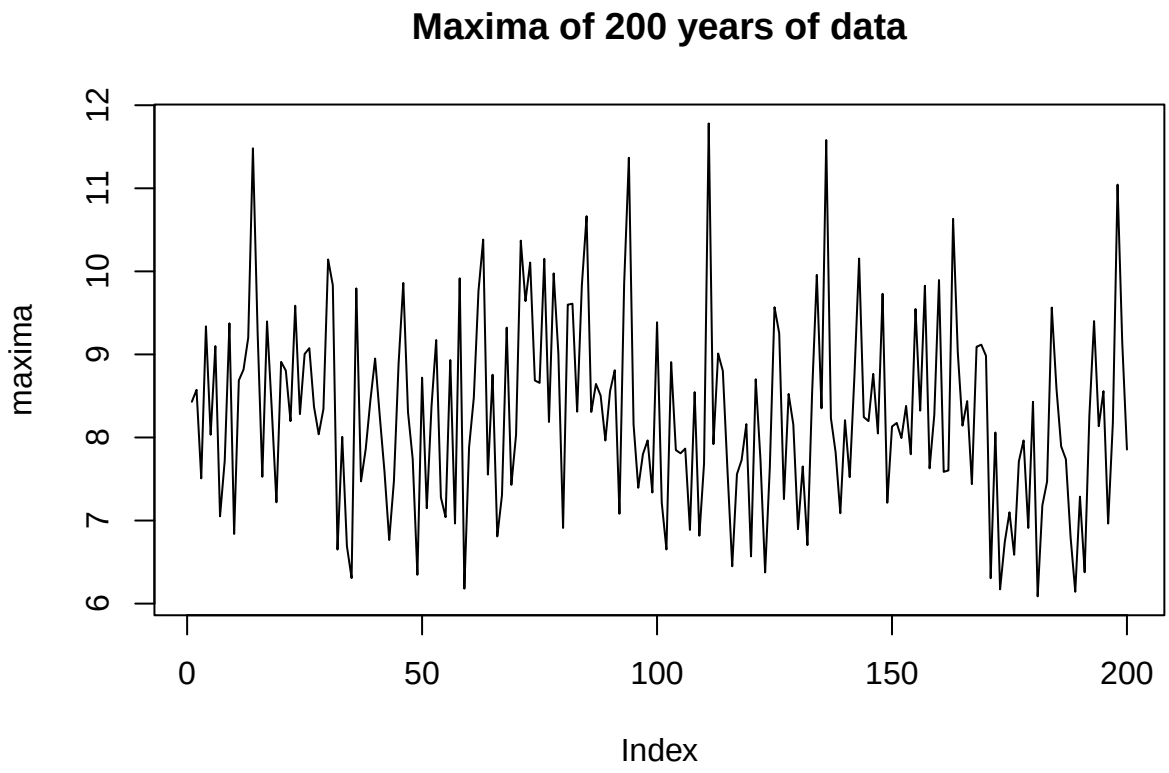



```
set.seed(123)

#block size
n<-12 #Assume monthly data for the year
original_mean<-5 #Assume the mean of the data is 5
original_sd<-2 #Assume the standard deviation of the data is 2

#generate monthly data over 200 years
series_length<-200
maxima<-c()
#Simulate 200 years of data
data_series<-replicate(series_length, rnorm(n=n,mean=original_mean,sd=original_sd), simplify = FALSE)
```

```r
#Check we have stored 200 years of data in blocks
print(data_series[1])
```

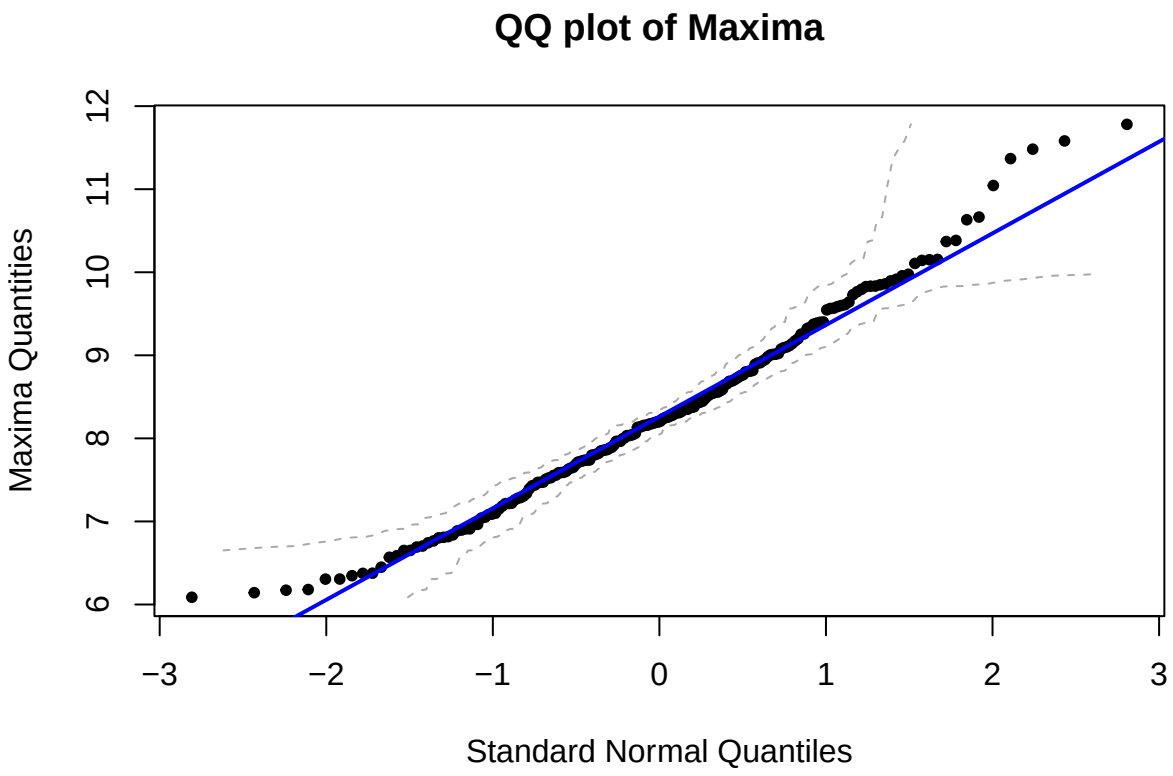
```
## [[1]]
##  [1] 3.879049 4.539645 8.117417 5.141017 5.258575 8.430130 5.921832 2.469878
##  [9] 3.626294 4.108676 7.448164 5.719628
```

```r
#Plot the data for the series
maxima<-unlist(lapply(data_series,FUN=max))
plot(maxima,main="Maxima of 200 years of data", type='l')
```



Lets evaluate the empirical distribution of the maxima in order to fit our data to a GEV distribution

```r
qqplot<-qqnorm(maxima, main='QQ plot of Maxima', ylab='Maxima Quantities')
qqline(maxima, col='blue', lwd=2)
```



Lets evaluate the empirical distribution of the maxima use Maximim Likelihood Estimation (MLE) using the fgev function in R to determine our parameters. This will give us an accur

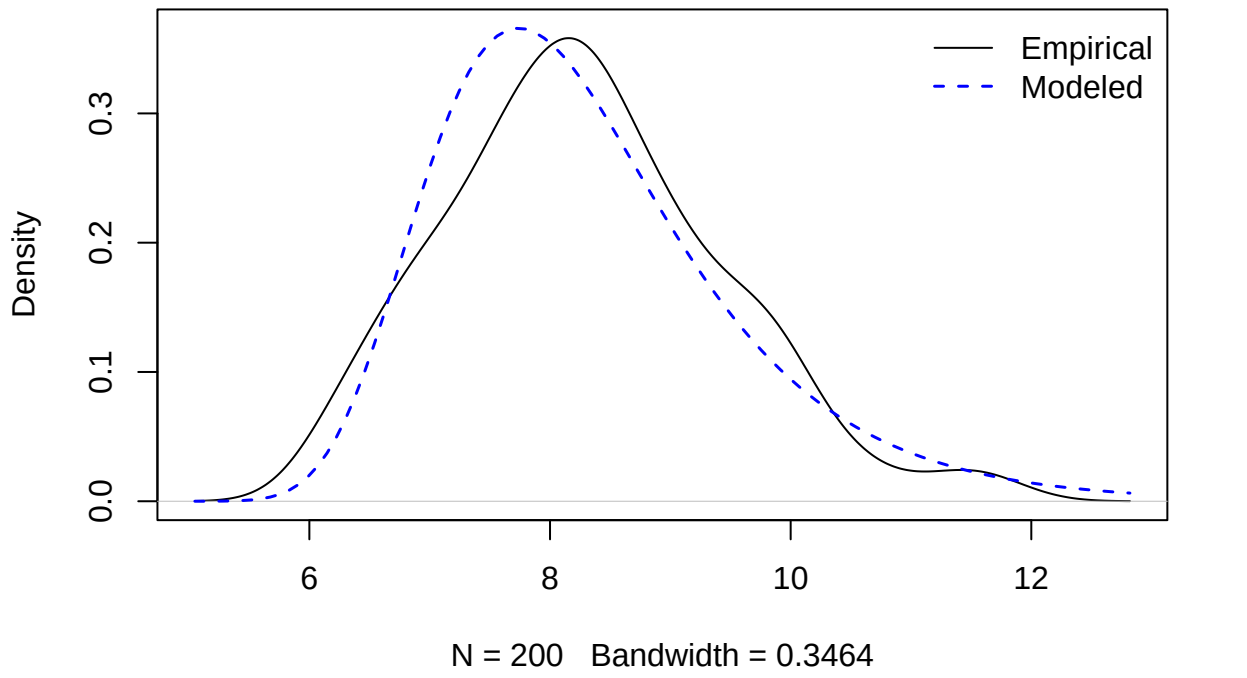
```
fit<-fgev(maxima)
gev_params<-fit$estimate
gev_params
```

```
##        loc      scale      shape
##  7.8167636  1.0393166 -0.1433933
```

```
location <- gev_params[1]
scale <- gev_params[2]
shape <- gev_params[3]
```

Lets fit a Gumbel distribution to the data as this will fit our data best and then plot the density of the data and our obtained shape parameter is close to 0 (the gumbel is a special case of GEV where the shape parameter $\epsilon \approx 0$)

```
fit_gum<-fevd(maxima, type='Gumbel')
plot(fit_gum,type='density', main='Empirical density evaluated against the estimated Gumbel distributio
```



Analyse the return level: what does it mean to say something is a 100 year event?

The return level plot shows the expected magnitude of extreme events over specific time intervals (e.g., once every 100 years). This is key for risk assessment, as it helps actuaries and analysts estimate how often and how severe rare, high-impact events might be.

Higher return levels for longer periods, like a 100-year return period, indicate potential for extreme events in the data. This information is valuable for predicting the risk of catastrophic events in fields like finance, environmental science, and insurance. Return levels give decision-makers insights into the likelihood of future extremes, helping them plan for rare but significant risks.

Take a quasntile $0 < p < 1$ and define

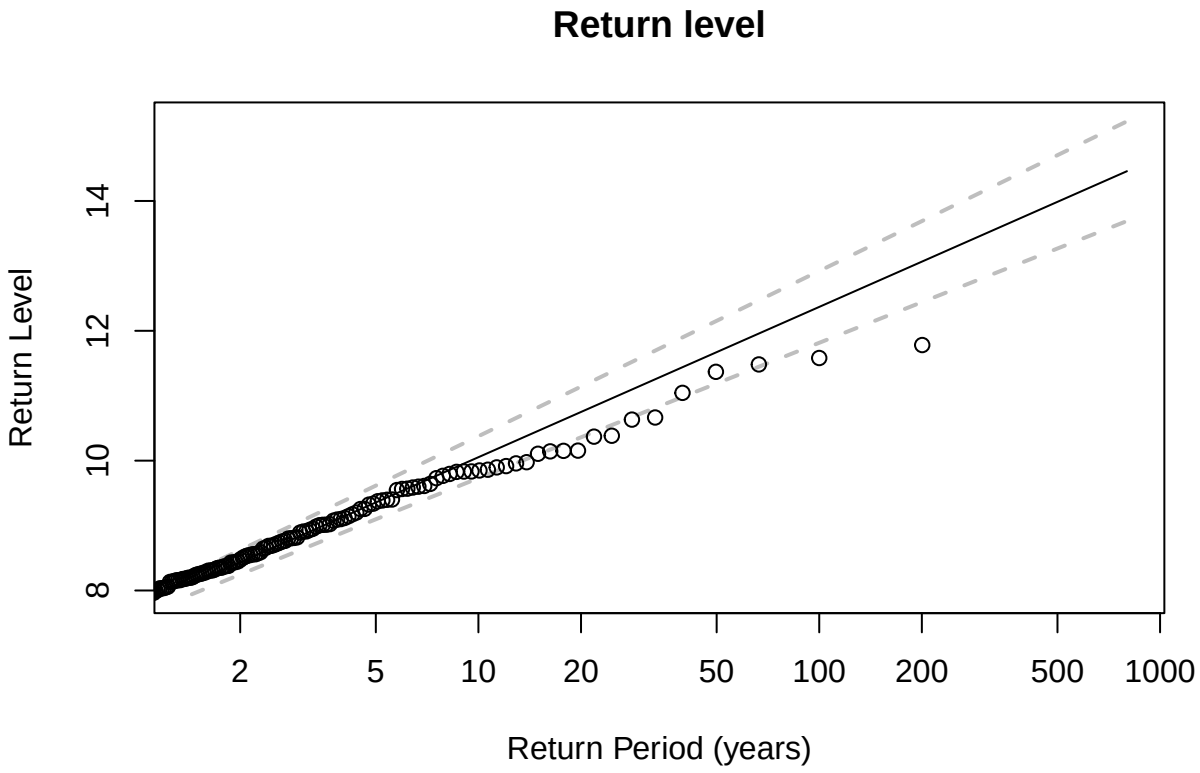
$$z_p = \mu - \frac{\sigma}{\xi}\left[1 - (-\log(1-p))^{-\xi}\right]$$

where $G(z_p) = 1 - p$

We now have a rainfall maxima level of z_p associated with a probability $1 - p$ and a return period of $1/p$ years.

The plot below gives

```
plot(fit_gum, type='rl', main = 'Return level')
```



Block Maxima Method Overview

Applications: Useful in fields requiring extreme event analysis, such as finance (e.g., extreme losses) and environmental science (e.g., floods and droughts).

Purpose: By dividing data into blocks and analyzing maxima, Block Maxima focuses on tail behavior, revealing rare, high-magnitude events.

Considerations: - **Block Size** m**:** Large m reduces maxima but may miss extremes; small m might not capture true extremes. - **Stationarity:** EVT assumes stationary data; non-stationary data may require detrending.

Challenges: - **Sample Size:** Small data sets affect GEV parameter reliability. - **Model Choice:** Selecting among Gumbel, Weibull, and Fréchet requires statistical tests or domain knowledge.

Peaks over Thresholds method

An alternative to Block Maxima, the Peaks Over Threshold method uses GPD to model values above a set threshold, efficient for frequent extremes.

Taking the random variable which will be our set of observations

$$X_1, X_2, X_3, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F$$

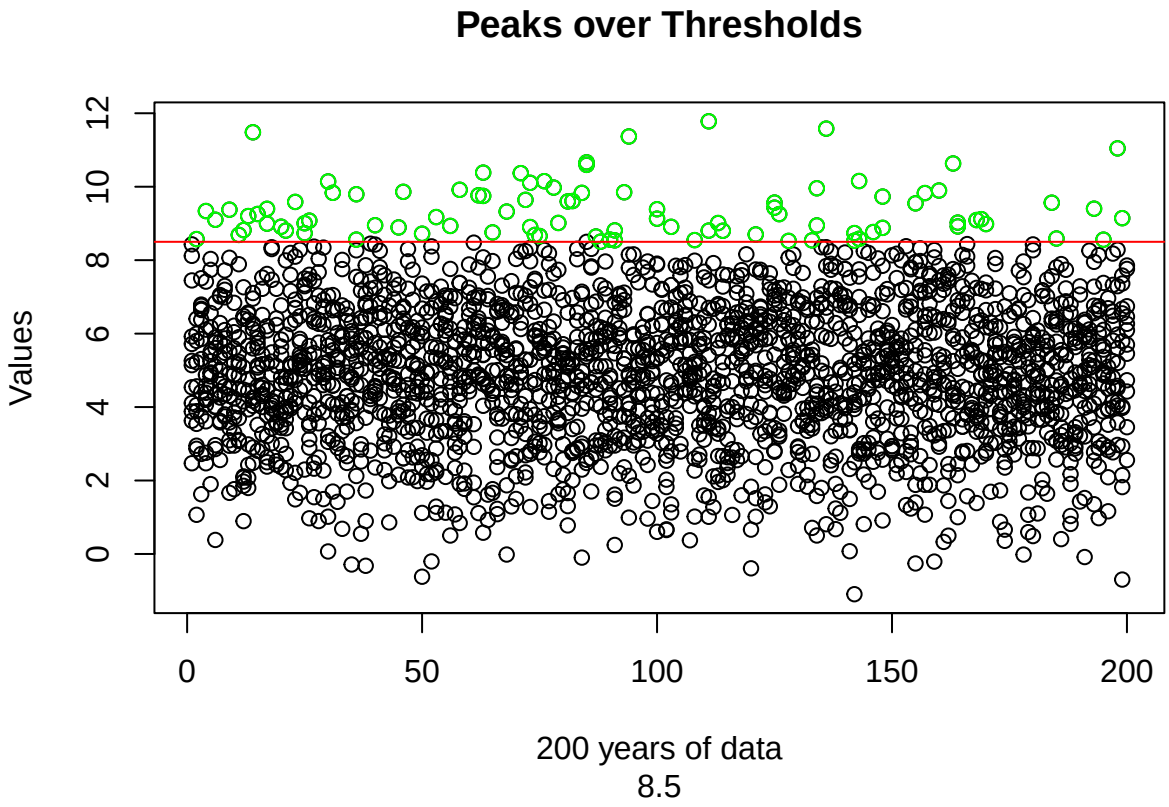
Our new variable Y would be the set of exceedences where $Y = X - u$ is defined for $X > u$ where u is the threshold of exceedence.

Using the indexing $Y_j = X_i - u$ to define the jth exceedence for $j = 1, ..., m_u$ we have

$$F(Y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X - u \leq x | X > u) Y \geq 0$$

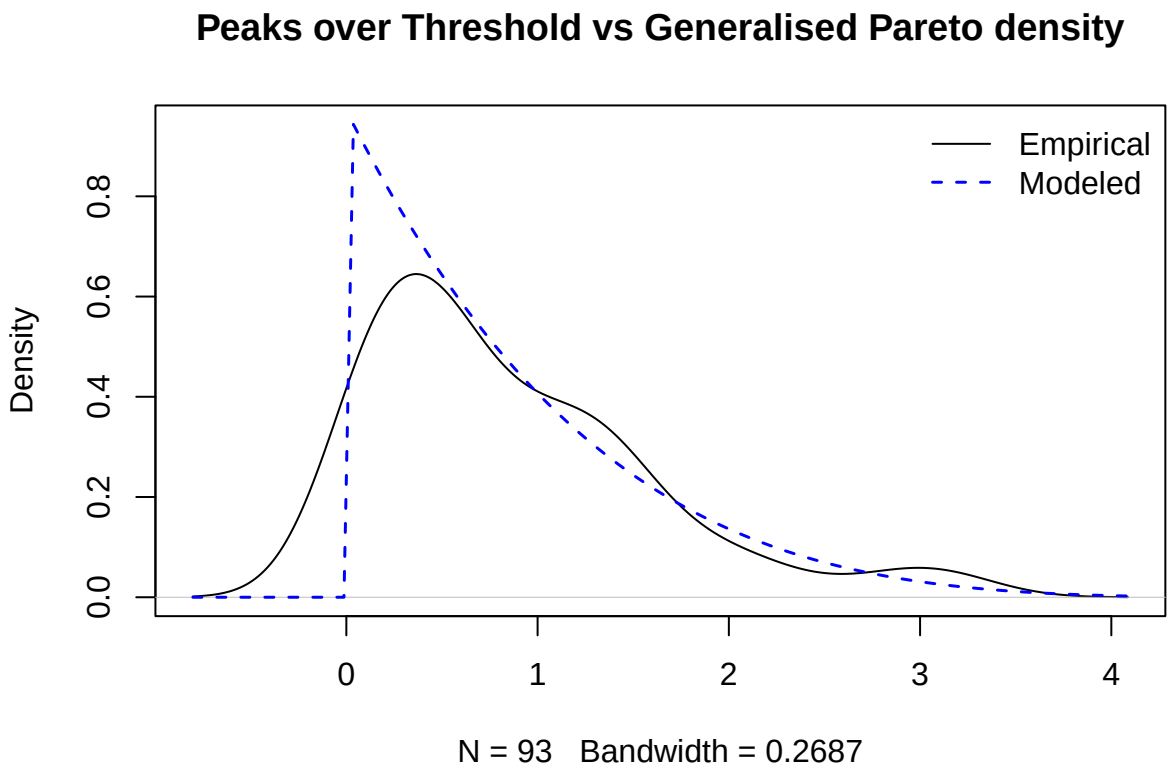
We will use a Generalised Pareto distribution to approximate this distribution as it is good for modelling the tails of a distribution.

```
#Set a theshold u to determine our peaks
threshold<-8.5
plot(x=rep(1:series_length, each=n), y = unlist((data_series)), main='Peaks over Thresholds', sub=paste
ylab='Values')
exceed_points<-which(unlist(data_series)>threshold)
points(x=rep(1:series_length, each=n)[exceed_points], y=unlist(data_series)[exceed_points], col='green'
abline(h=threshold, col='red')
```



Lets fit a Generalised Pareto distribution to the data and plot the density of the data to visualise the fit

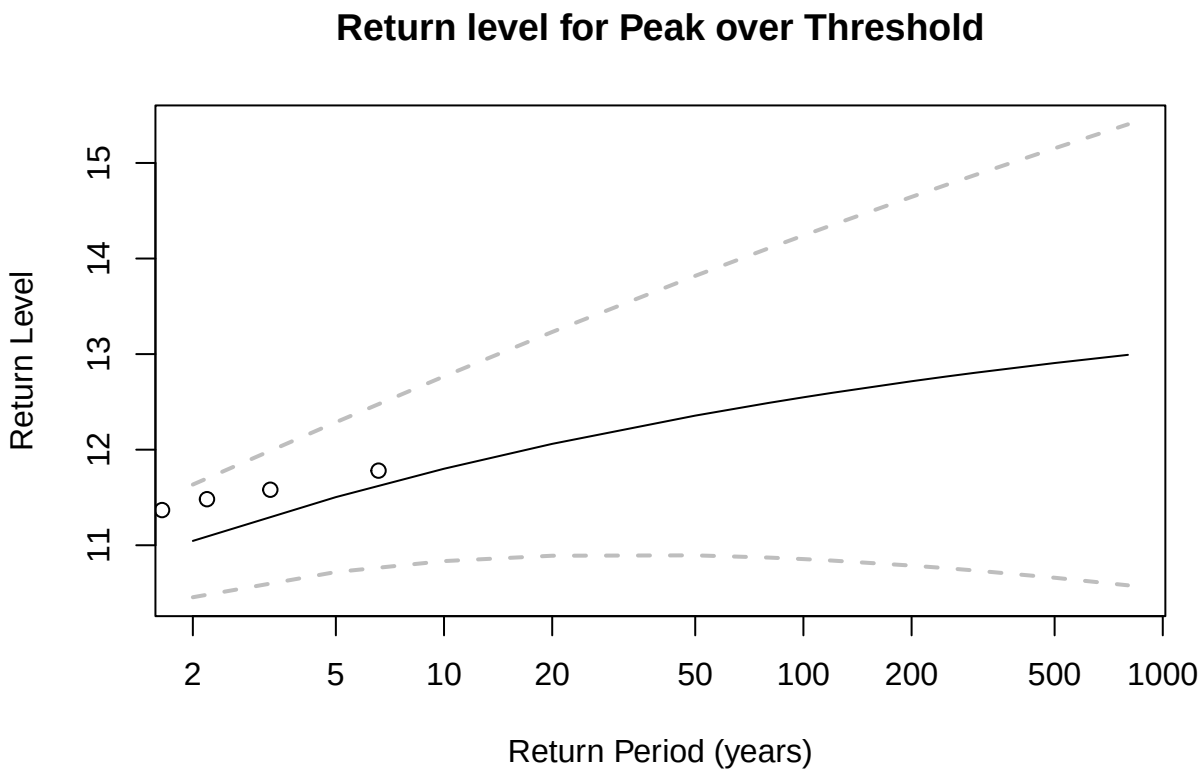
```
gen_par_fit<-fevd(unlist(data_series), threshold = threshold, type = 'GP')
scale <- gen_par_fit$estimate[1]
shape <- gen_par_fit$estimate[2]
plot(gen_par_fit, type='density', main = 'Peaks over Threshold vs Generalised Pareto density')
```



The fitted GPD describes the distribution of values that exceed the threshold. A good fit indicates that the data has heavy tails, meaning extreme events are relatively likely, and provides an estimate of how severe these extreme values can be. This is especially valuable in insurance, where understanding the likelihood and impact of rare events is crucial for managing risk.

Now lets evaluate the return level for the peaks over threshold method

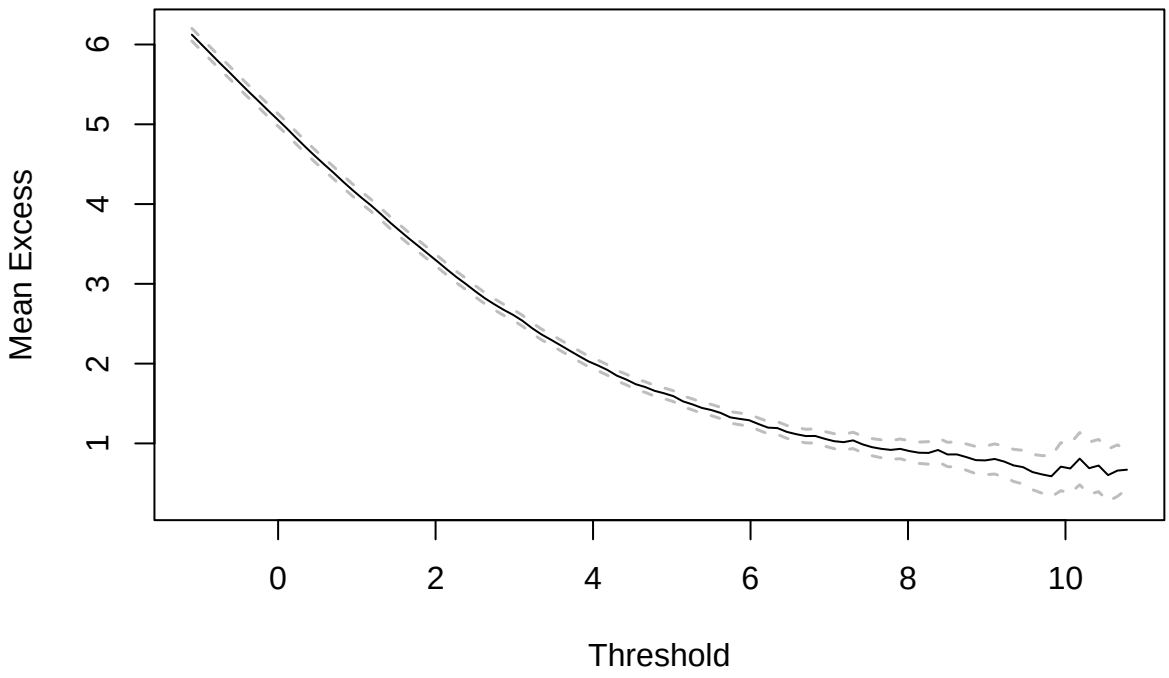
```
plot(gen_par_fit, type ='rl', main='Return level for Peak over Threshold')
```



Optimisation for Model

We picked a fairly arbitrary threshold of 8.5. Instead we can use the Mean Residual Life(MRl) plot to aid in the selection of a threshold. Look for a flat section of the curve, the thresholds within this region tend to be good candidates indicating that the data above these threshholds follows a GPD. Choose a point of stability, often the lowest threshold associated with the flat region which balances a sufficient number of exceedances with a stable GPD fit.

```
# Mean Residual Life Plot
extRemes::mrlplot(unlist(data_series), threshold.range = seq(5, 10, by = 0.5))
```



Use a goodness-of-fit test to assess wether a GDP is appropriate for our data series

```
gpd_model <- fpot(unlist(data_series), threshold)
gpd_model
```

```
##
## Call: fpot(x = unlist(data_series), threshold = threshold)
## Deviance: 155.9578
##
## Threshold: 8.5
## Number Above: 93
## Proportion Above: 0.0388
##
## Estimates
##   scale    shape
##  1.0293  -0.1904
##
## Standard Errors
##  scale   shape
## 0.1478  0.1013
##
## Optimization Information
##   Convergence: successful
##   Function Evaluations: 38
##   Gradient Evaluations: 9
```

Look at the confidence intervals for the parameters of the GPD using profile-liklihood based confidence intervals

```
ci(gen_par_fit)
```

```
## fevd(x = unlist(data_series), threshold = threshold, type = "GP")
##
## [1] "Normal Approx."
##
## [1] "100-year return level: 12.548"
##
## [1] "95% Confidence Interval: (10.8553, 14.2409)"
```

For time-varying return levels, using non-stationary GPD models where parameters are able to change over time allows you to analyze how return levels may shift with trends, like climate change effects. Usd multiple thresholds if your data exhibits a complex tail structure. This can involve fitting different GPDs to different parts of the tail, which is useful for analyzing data with varying degrees of desired outliers.

Point Process Representation

Consider a series of i.i.d random variables $X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F$ for some unknown distribution F. Construct a point process

$$(X_i, i); i = 1, 2, \ldots, n$$

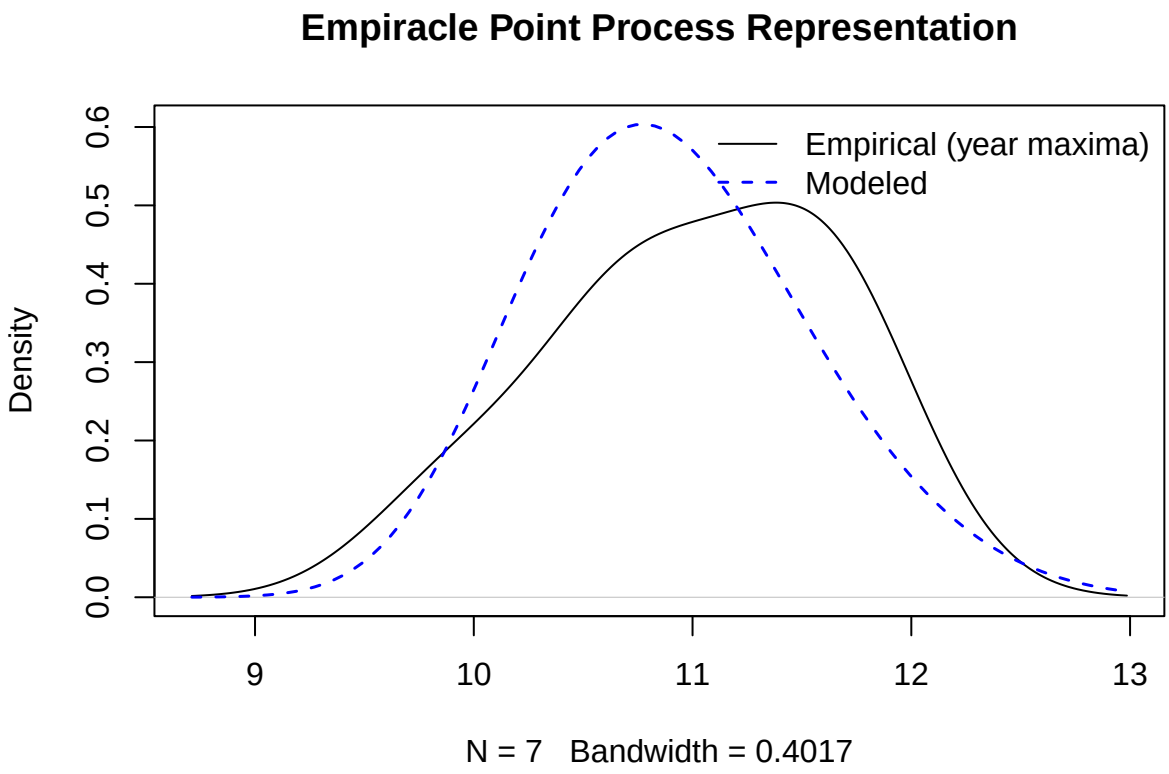
and model the the behaviour of this process in 2-dimensions as points that follow: - i in some interval $[t_1, t_2]$ where $t_1 < t_2$,
- X_i in an interval $u \leq X_i \leq \infty$ for some threshold u.

We then have a region $A = [t_1, t_2] \times [u, \infty)$ containing our points.Our sequence of point processes on $\mathbb{R}^2$ is then defined as

$$\mathbb{P}_n = \left\{ \left(\frac{i}{n+1}, \frac{X_i - b_n}{a_n} \right) : i = 1, \ldots, n \right\}$$

where $P_N \to P$ as $n \to \infty$ where P is a Poisson process with intensity λ

```
point_fit<-fevd(unlist(data_series), threshold = threshold, type ='PP')
plot(point_fit, type = 'density', main = 'Empiracle Point Process Representation')
```


Empiracle Point Process Representation
Empirical (year maxima)
Modeled
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
9
10
11
12
13
N = 7 Bandwidth = 0.4017