MATH5805 – A Modern Introduction to Extreme Value Theory

Week 3 - Univariate Extremes: the Block Maxima approach (II)

Dr Boris Béranger
School of Mathematics and Statistics
UNSW Sydney

CRICOS Provider Code 00098G

Lecture outline

1. Introduction
2. Return Levels and Return Periods
3. Parameter Estimation
4. Analysis of Precipitation Data

CRICOS Provider Code 00098G

1. Introduction



Goals



Week 2 & 3 are about introducing the **Block Maxima** approach.

This week, we

- Define the concepts of **return levels** and **return periods**, which are commonly used to derive information about future extreme events.
- Discuss the **estimation** of the parameters of the GEV distribution.
- Use the block maxima approach to **model a precipitation dataset**.

Note that throughout this lecture we will make the assumption that $X_1, \ldots, X_n$ is a sequence of i.i.d. random variables with **common distribution function** F.



2. Return Levels and Return Periods

Return Levels and Return Periods

Definition

A **return level** is a high quantile for which the variable of interest (maximum) has probability $1/$**return period** to exceed. Mathematically this translate to

$$\mathbb{P}(M_n > \text{return level}) = \frac{1}{\text{return period}}.$$

The above is equivalent to

$$\mathbb{P}(M_n \leq \text{return level}) = 1 - \frac{1}{\text{return period}} \iff \text{return level} = U(\text{return period}).$$

Week 2 lecture: we defined $U(n)$ as the return level for the return period n.

Example

The 100-year flood/rainfall event is the high quantile for which the probability of the rainfall annual maximum exceeding this quantile is $1/100$.

Return Levels and Return Periods

Using the Extremal Type Theorem and GEV limiting distribution, an estimate of the return level is

$$z_{\mathbf{p}} = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - \{-\log(1-\mathbf{p})\}^{-\xi}\right] & \xi \neq 0, \\ \mu - \sigma \log\left(-\log\left(1-\mathbf{p}\right)\right) & \xi = 0, \end{cases} \tag{1}$$

where $\mathbf{p} = 1/\text{return period}$. Setting $y_p = -\log(1-p)$ we have

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

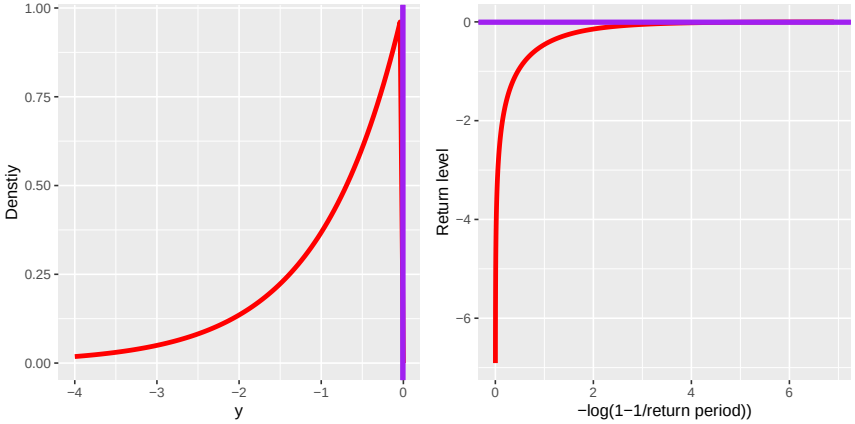
Hence if we plot z_p against $\log(y_p)$ then:

- if $\xi = 0$, the plot is **linear**;
- if $\xi < 0$, the plot is **convex** with asymptotic limit $\mu - \sigma/\xi$ as $p \to 0$;
- if $\xi > 0$, the plot is **concave** with no finite bound;

Return Levels and Return Periods

$$\theta = (-1, 1, -1) \quad \text{[Weibull]}$$

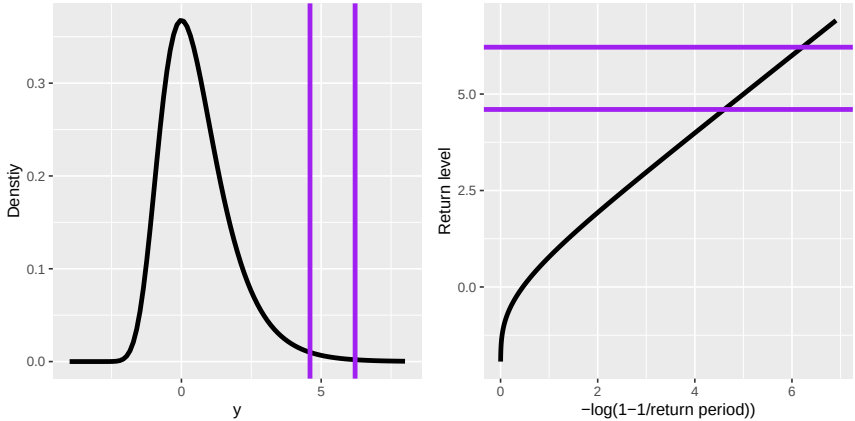
Return levels for return period: 100, 500



Return Levels and Return Periods

$\theta = (0, 1, 0)$ [Gumbel]

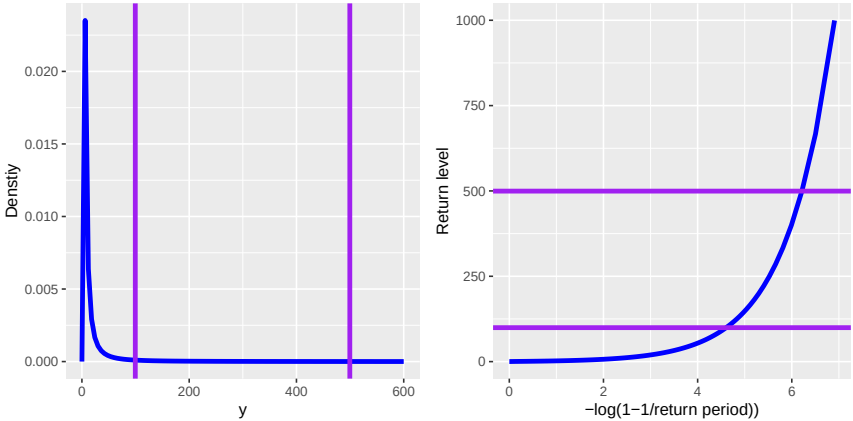
Return levels for return period: 100, 500



Return Levels and Return Periods

$\theta = (1, 1, 1)$ **[Fréchet]**

Return levels for return period: 100, 500





3. Parameter Estimation

Parameter Estimation

Block maxima approach assumes that maxima taken from independent blocks are available, and that they can be fitted with a GEV distribution.

- **Larger blocks** mean **fewer maxima** and **larger variance**
- **Smaller blocks** may not fit the distribution and lead to **bias**

Often block size is determined from the **pragmatic perspective** or **data availability**, leading to data such as annual, monthly or daily maxima.

Suppose that $Z_1, \ldots, Z_m$ are maxima from m independent blocks and assumed to be from the GEV distribution.

Parameter estimation

There is **one issue** with the maximum likelihood estimator (MLE). Recall that the support of **the GEV distribution depends on the parameter values**:

- $\mu - \sigma/\xi$ is the upper endpoint for $\xi < 0$ and lower endpoint for $\xi > 0$;
- the support is unbounded for $\xi = 0$.

$\Longrightarrow$ **Regularity conditions for the asymptotic properties of the MLE do not hold**.

Parameter estimation

Result [Smith (1985)]

- When $\xi > -0.5$, the MLE has the usual asymptotic properties;
- When $-1 < \xi \leq -0.5$, the MLE exists but does not have that standard asymptotic properties (especially for μ);
- When $\xi \leq -1$, the MLE does not exist.

The case $\xi \leq -0.5$ correspond to distributions with bounded upper endpoint, which is not as frequently observed as the cases with $\xi > 0.5$ in practice.

Maximum likelihood estimation

When $\xi \neq 0$, the log-likelihood function is given by

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^{m} \log g(y_i; \theta) \\ &= -m \log \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{m} \log \left(1 + \xi \frac{y_i - \mu}{\sigma}\right) \quad - \sum_{i=1}^{m} \left(1 + \xi \frac{y_i - \mu}{\sigma}\right)^{-1/\xi},\end{aligned}$$

provided that

$$1 + \xi \left(\frac{y_i - \mu}{\sigma}\right) > 0, \text{for } i = 1, \ldots, m. \tag{2}$$

When $\xi = 0$, the likelihood is given by

$$\ell(\theta) = -m \log \sigma - \sum_{i=1}^{m} \left(\frac{z_i - \mu}{\sigma}\right) - \sum_{i=1}^{m} \exp\left\{-\left(\frac{z_i - \mu}{\sigma}\right)\right\}.$$

Maximum likelihood estimation

Maximisation of the likelihood function with respect to $\theta = (\mu, \sigma, \xi)$ leads to MLE with respect to the entire GEV family.

There is no analytical solution but it is a straight forward using **numerical optimisation procedures** (need to ensure (2) holds).

The **score** of the likelihood and **Fisher information** have been derived by **Prescott and Walden (1980)**.

Maximum likelihood estimation



Exercise

1. Write down the derivative of the log-likelihood necessary to derive the scores.
2. Derive the expression of the entries of the Fisher information matrix.

Solutions

The derivatives of the log-density are:

$$\frac{\partial \log g}{\partial \mu} = \frac{1}{\sigma w}(1+\xi-v),$$
$$\frac{\partial \log g}{\partial \sigma} = \frac{1}{\sigma}\left[-1+\frac{1}{\sigma w}(1+\xi-v)(y-\mu)\right],$$
$$\frac{\partial \log g}{\partial \xi} = \frac{1}{\xi^2}\log(w)(1-v) - \frac{1}{\sigma w}(1+\xi-v)(y-\mu)\xi,$$

where $w = 1+\xi(y-\mu)/\sigma$ and $v = w^{-1/\xi}$.
The Fisher information is $m\mathcal{I}$ where $\mathcal{I}$ has entries:

$$\mathcal{I}_{\mu\mu} = \frac{p}{\sigma^2},$$
$$\mathcal{I}_{\sigma\sigma} = \frac{1}{\sigma^2\xi^2}\left[1-2\Gamma(2+\xi)+p\right],$$

Maximum likelihood estimation

$$\mathcal{I}_{\xi\xi} = \frac{1}{\xi^2}\left[\frac{\pi^2}{6} + \left(1 - \gamma + \frac{1}{\xi}\right)^2 - \frac{2q}{\xi} + \frac{p}{\xi^2}\right],$$
$$\mathcal{I}_{\mu\sigma} = -\frac{1}{\sigma^2\xi}\left[p - \gamma(2+\xi)\right],$$
$$\mathcal{I}_{\mu\xi} = \frac{1}{\sigma\xi}\left(q - \frac{p}{\xi}\right),$$
$$\mathcal{I}_{\sigma\xi} = \frac{1}{\sigma\xi^2}\left[1 - \gamma + \frac{1-\Gamma(2+\xi)}{\xi} - q + \frac{p}{\xi}\right],$$

where $p = (1+\xi)^2\,\Gamma(1+2\xi)$, $q = \Gamma(2+\xi)\left[\Gamma'(1+\xi) + \frac{1}{\xi} + 1\right]$, $\Gamma'(r) = \mathrm{d}\log\Gamma(r)\mathrm{d}r$ is the digamma function and γ is Euler's constant $-\Gamma'(1)$.
Note that the information is only valid for $\xi > -0.5$.

Maximum likelihood estimation

The MLE $\hat{\theta}$ of the GEV parameters θ is asymptotically consistent, normally distributed, and most efficient when $\xi > -0.5$.

Confidence intervals and other forms of inference follow immediately from the approximate normality of the MLE.

For small-to moderate sample sizes, **the MLE is not robust**:

- can be **hard to find** (numerical issues);
- can have **large variance** compared to the probability weighted moments approach (not covered in this course).

Inference for returns levels

Return level estimation is often of primary interest.

Using the definition of return levels given by (1) a return level is estimated as

$$\hat{z}_p = \begin{cases} \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\left[1 - y_p^{-\hat{\xi}}\right] & \hat{\xi} \neq 0, \\ \hat{\mu} - \hat{\sigma}\log\left(y_p\right) & \hat{\xi} = 0, \end{cases}$$

Confidence intervals can be obtained in principle with the **delta method** using V, the asymptotic variance of $\hat{\theta}$:

$$\mathrm{Var}(\hat{z}_p) \approx \Delta z_p^\top V \Delta z_p,$$

with

$$\Delta z_p^\top = \left[\frac{\partial z_p}{\partial \mu}, \frac{\partial z_p}{\partial \sigma}, \frac{\partial z_p}{\partial \xi}\right] = \left[1, \frac{-(1-y_p^{-\xi})}{\xi}, \frac{\sigma(1-y_p^{-\xi})}{\xi^2} - \frac{\sigma \log y_p}{\xi y_p^{\xi}}\right],$$

evaluated at $\hat{\theta}$.

Inference for returns levels

Usually the interest is in **long return periods** with **small probability** p.


Performance might be poor, especially for long return periods.

- Normal approximation to the distribution of the MLE might be poor

$\Longrightarrow$ **profile likelihood**

- Estimates and their measure of precision are based on **the assumption that the model is correct**

$\Longrightarrow$ extrapolation is based on unverifiable assumptions, and measures of uncertainty on returns levels should be regarded as lower bounds that could be much greater if uncertainty due to model correctness is taken into account.

Profile likelihood

The **profile likelihood** approach considers **partionning θ into (θ_1, θ_2)** where θ_1 contains a subvector of dimension k of current interest. The profile log-likelihood for θ_1 is

$$\ell_{\text{prof}}(\theta_1) = \max_{\theta_2} \ell((\theta_1, \theta_2)).$$

Theorem [C01, Theorem 2.6]

Under suitable regularity conditions, the deviance

$$D_{\text{prof}}(\theta_1) = 2\left[\ell(\hat{\theta}) - \ell_{\text{prof}}(\theta_1)\right] \sim \chi^2_k.$$

This theorem can be used to construct $1-\alpha$ confidence region for θ_1 with $C_\alpha = \{\theta_1 : D_{\text{prof}}(\theta_1) \leq c_\alpha\}$, where c_α is the α upper quantile of χ^2_k.

Profile likelihood

Example Take $\theta_1 = \xi_0$.

- For a fixed $\xi = \xi_0$, we maximise the log-likelihood w.r.t μ and σ;
- The procedure is repeated for a range of values of ξ_0;
- The corresponding maximised values of the log-likelihood constitute the profile log-likelihood for ξ.

For **return levels** one can re-parametrise the GEV distribution in terms of z_p.



Exercise

Write down the re-parametrised GEV distribution in terms of z_p.

4. Analysis of Precipitation Data



CRICOS Provider Code 00098G

Analysis of precipitation data

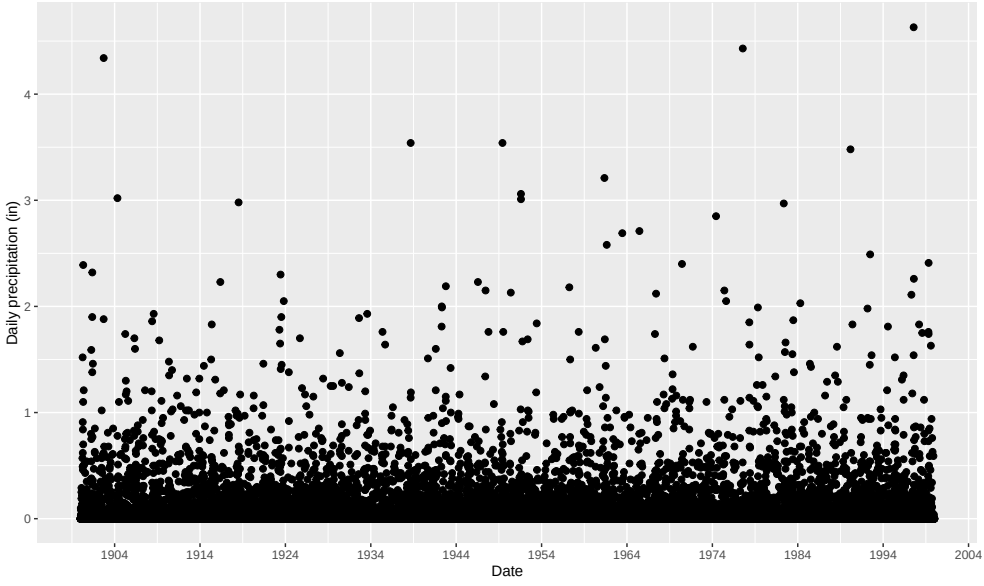
We consider the **Fort Collins, Colorado daily precipitation amounts**

- Times series of daily precipitation amount (in) between 1900 and 1999.
- Semi-arid region.
- Marked annual cycle in precipitation (wettest in late spring/early summer, driest in winter).
- No obvious long term trend.
- **28 July 1997 flood**: caused substantial damage to Colorado State University.

Available in the `R` package `extRemes`. See **Katz et al. (2002)** for more details.

```
library(extRemes)
data(Fort)
```

Daily precipitation amount in Fort Collins, Colorado



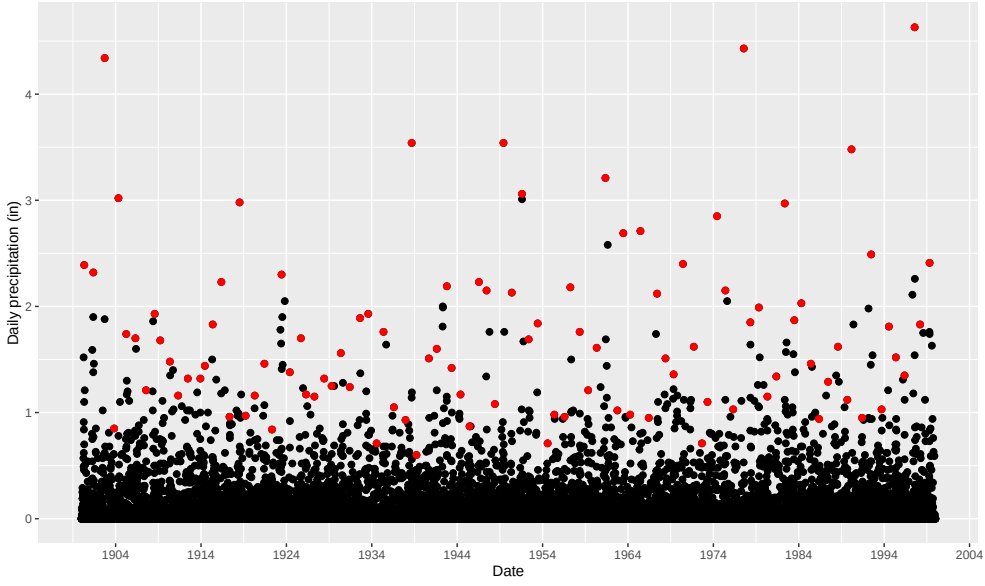
Daily precipitation amount in Fort Collins, Colorado

Code to reproduce the previous plot:

```
Fort$date <- as.Date(with(Fort, paste(year, month, day,sep="-")),
                     "%Y-%m-%d")

ggplot(Fort, aes(date,Prec)) +
  geom_point(size=2) +
  scale_x_date(date_breaks = "10 year", date_labels =  "%Y") +
  xlab("Date") +
  ylab("Daily precipitation (in)")
```

Daily precipitation amount in Fort Collins, Colorado + Annual maxima



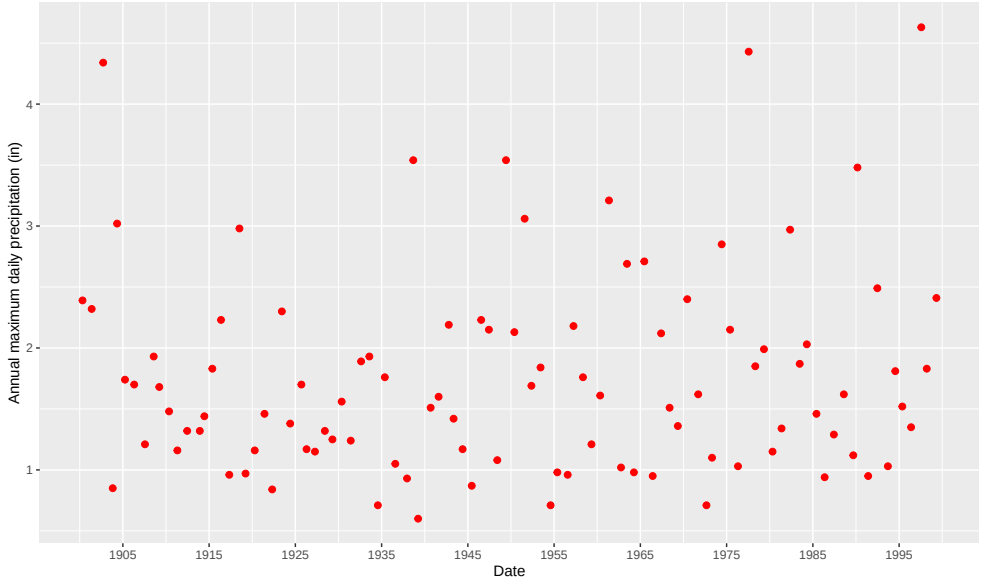
Daily precipitation amount in Fort Collins, Colorado + Annual maxima

Code to reproduce the previous plot:

```
Fort.bm <- blockmaxxer(Fort, blocks = Fort$year, which="Prec")

ggplot(Fort, aes(date,Prec)) +
  geom_point(size=2) +
  scale_x_date(date_breaks = "10 year", date_labels =  "%Y") +
  xlab("Date") +
  ylab("Daily precipitation (in)") +
  geom_point(data = Fort.bm, mapping = aes(date, Prec),
             colour="red", size=2)
```

Daily precipitation amount in Fort Collins, Colorado, Annual maxima only



Daily precipitation amount in Fort Collins, Colorado, Annual maxima only

```
ggplot(Fort.bm, aes(date,Prec)) +
  geom_point(size=2, col="red") +
  scale_x_date(date_breaks = "10 year", date_labels =  "%Y") +
  xlab("Date") +
  ylab("Annual maximum daily precipitation (in)")
```

Analysis of precipitation data

How often is an extreme such as the July 1997 event expected?

- Assuming no long-term trends emerges.
- Using annual maxima removes effects of seasonal trend in analysis.

Step 1: We start by fitting the Gumbel case ($\xi = 0$)

```
fit.gum <- fevd(Prec, Fort.bm, type="Gumbel", units="in")
# fit.gum
```

Analysis of precipitation data

```r
ci(fit.gum, type="parameter")
```

```
## fevd(x = Prec, data = Fort.bm, type = "Gumbel", units = "in")
##
## [1] "Normal Approx."
##
##          95% lower CI  Estimate 95% upper CI
## location    1.2800537 1.3988265    1.5175994
## scale       0.4856291 0.5784564    0.6712836
```

Analysis of precipitation data

The estimated parameters are

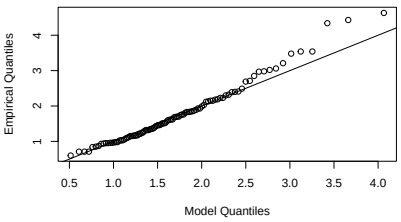
Parameter	Estimate	Std. Error	Conf. Int.
Location (μ)	1.399	0.061	$(1.280, 1.518)$
Scale (σ)	0.578	0.047	$(0.486, 0.671)$

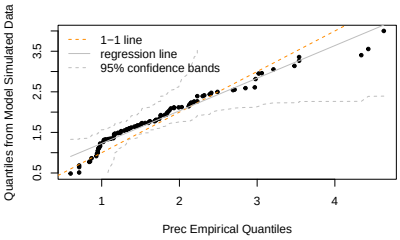
We assess the quality of the fit by running some diagnostic plots

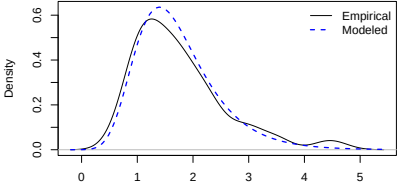
```
plot(fit.gum)
plot(fit.gum,"trace")
```

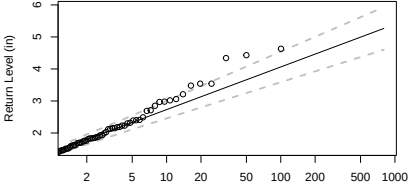
Analysis of precipitation data

fevd(x = Prec, data = Fort.bm, type = "Gumbel", units = "in")

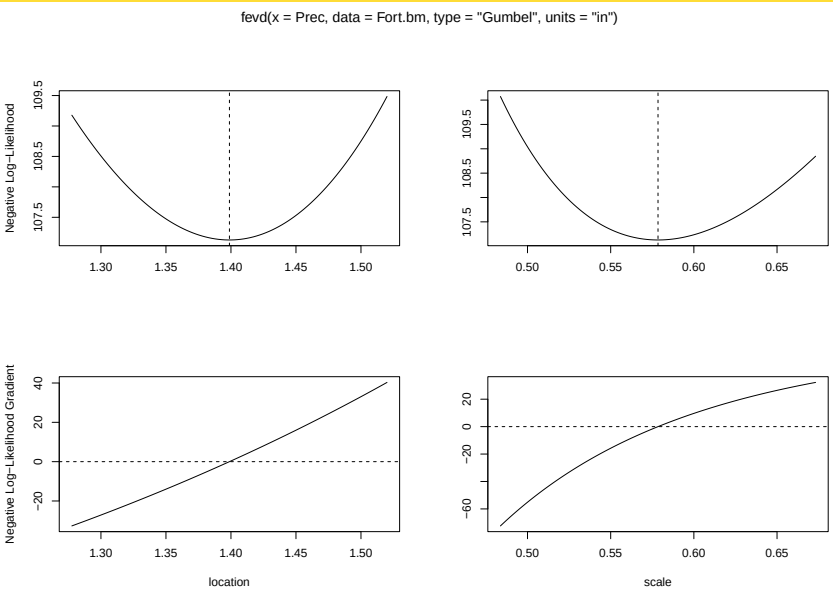








Analysis of precipitation data



Analysis of precipitation data

The return levels associated with return periods of 20, 50 and 100 years are

Return period	Estimate	Conf. Int.
20-year	3.117	(2.786, 3.448)
50-year	3.656	(3.242, 4.069)
100-year	4.060	(3.584, 4.536)

```
return.level(fit.gum, return.period=c(20,50,100), do.ci=TRUE)
```

```
## fevd(x = Prec, data = Fort.bm, type = "Gumbel", units = "in")
##
## [1] "Normal Approx."
##
##                        95% lower CI Estimate 95% upper CI
## 20-year return level       2.785827 3.116955     3.448082
## 50-year return level       3.242475 3.655928     4.069381
## 100-year return level      3.583699 4.059812     4.535925
```

Step 2: We now fit the GEV

```
fit.gev <- fevd(Prec, Fort.bm, units="in")
#fit.gev
ci(fit.gev, type="parameter")
```

```
## fevd(x = Prec, data = Fort.bm, units = "in")
##
## [1] "Normal Approx."
##
##          95% lower CI  Estimate 95% upper CI
## location  1.225753534 1.3466597    1.4675658
## scale     0.437181061 0.5328046    0.6284282
## shape    -0.006601292 0.1736264    0.3538540
```

Analysis of precipitation data

The estimated parameters are

Parameter	Estimate	Std. Error	Conf. Int.
Location (μ)	1.347	0.062	$(1.226, 1.468)$
Scale (σ)	0.533	0.049	$(0.437, 0.628)$
Shape(ξ)	0.174	0.092	$(-0.07, 0.354)$

Seems to be heavy tailed!

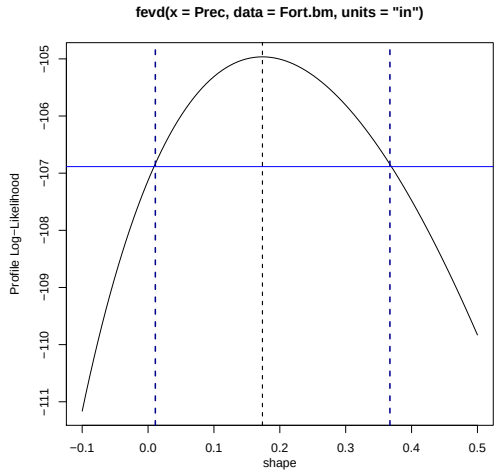
Let's have a closer look at the shape parameter ξ

```
ci(fit.gev, type="parameter", which.par=3, xrange=c(-0.1,0.5),
   nint=100, method="proflik", verbose=TRUE)
```

Analysis of precipitation data

```
 ## Preparing to calculate 95 % CI for shape parameter
## Model is fixed
## Using Profile Likelihood Method.
##
## Calculating profile likelihood.  This may take a few moments.
##
## Profile likelihood has been calculated.  Now, trying to find where
## it crosses the critical value = -106.8853
## fevd(x = Prec, data = Fort.bm, units = "in")
##
## [1] "Profile Likelihood"
## [1] "shape:  0.174"
## [1] "95% Confidence Interval:  (0.011, 0.3671)"
```

Analysis of precipitation data



Using the profile likelihood confirms that there is a heavy tail phenomena.

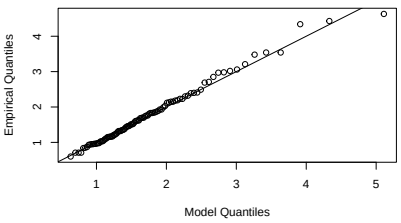
Analysis of precipitation data

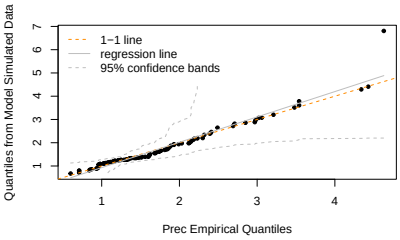
Again, we assess the quality of the fit by running some diagnostic plots

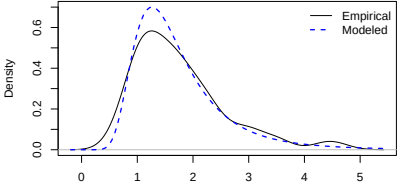
```
plot(fit.gev)
plot(fit.gev,"trace")
```

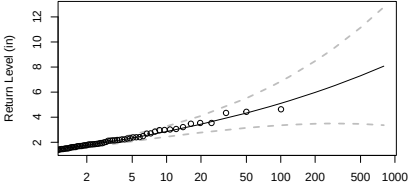
Analysis of precipitation data

fevd(x = Prec, data = Fort.bm, units = "in")

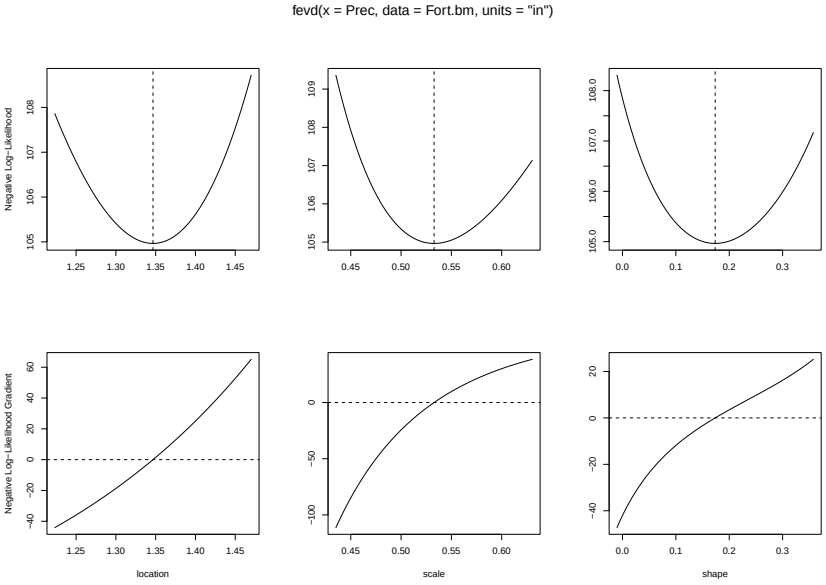








Analysis of precipitation data



Analysis of precipitation data

The return levels associated with return periods of 20, 50 and 100 years are

Return period	Estimate	Conf. Int.
20-year	3.417	$(2.765, 4.070)$
50-year	4.320	$(3.145, 5.495)$
100-year	5.100	$(3.354, 6.843)$

Analysis of precipitation data

The numbers were obtained from:

```
return.level(fit.gev, return.period=c(20,50,100), do.ci=TRUE)
```

```
## fevd(x = Prec, data = Fort.bm, units = "in")
##
## [1] "Normal Approx."
##
##                         95% lower CI Estimate 95% upper CI
## 20-year return level        2.765057 3.417462     4.069868
## 50-year return level        3.144981 4.319935     5.494890
## 100-year return level       3.354204 5.098635     6.843067
```

Step 3: We test if $\xi = 0$

We perform a likelihood ratio test: $H_0 : \xi = 0$ (Gumbel type) against $H_1 : \xi > 0$

```
lr.test(fit.gum, fit.gev)
```

```
##
##  Likelihood-ratio Test
##
## data:  PrecPrec
## Likelihood-ratio = 4.3264, chi-square critical value = 3.8415, alpha =
## 0.0500, Degrees of Freedom = 1.0000, p-value = 0.03752
## alternative hypothesis: greater
```

Since p-value ≈ 0.038 **we reject the null hypothesis** (this was expected based on the CI for ξ).

Step 4: What is the probability of an annual maximum precipitation to be at least as large as that during the 28 July 1997 flood (4.63 in)?

We are interested in

$$\mathbb{P}(M_n \geq 4.63).$$

```
fit.gev.sum <- summary(fit.gev, silent=TRUE)
pevd(4.63, loc=fit.gev.sum$par[1], scale=fit.gev.sum$par[2],
     shape=fit.gev.sum$par[3], lower.tail=FALSE)
```

```
## [1] 0.01502957
```

References



Sources and recommending reading

1. **[dHF06]** de Haan, L. & Ferreira, A. (2006) Extreme Value Theory: An Introduction. New York: Springer.
2. **[C01]** Coles, S. (2001) An Introduction to Statistical Modeling of Extreme Values. Springer. Chapter 1-3.
3. **Katz, R. W., Parlange, M. B. and Naveau, P. (2002)**. Statistics of extremes in hydrology, *Advances in Water Resources*, 25, pp.1287–1304.