UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

A Estatística da Confiabilidade

Análise de Valores Extremos

Felipe Campelo
http://www.cpdee.ufmg.br/~fcampelo

Departamento de Engenharia Elétrica

Belo Horizonte
Agosto de 2013

Introdução

Motivação

A modelagem e inferência estatística usualmente discutidas tratam principalmente do *comportamento médio* de sistemas, processos, fenômenos ou dispositivos;

Entretanto, em análise de confiabilidade o comportamento dos *extremos* é frequentemente mais importante:

- Máxima velocidade de vento esperada para uma dada região;
- Mínima resistência de dados materiais;
- Pior desempenho esperado de determinado componente em um sistema complexo.

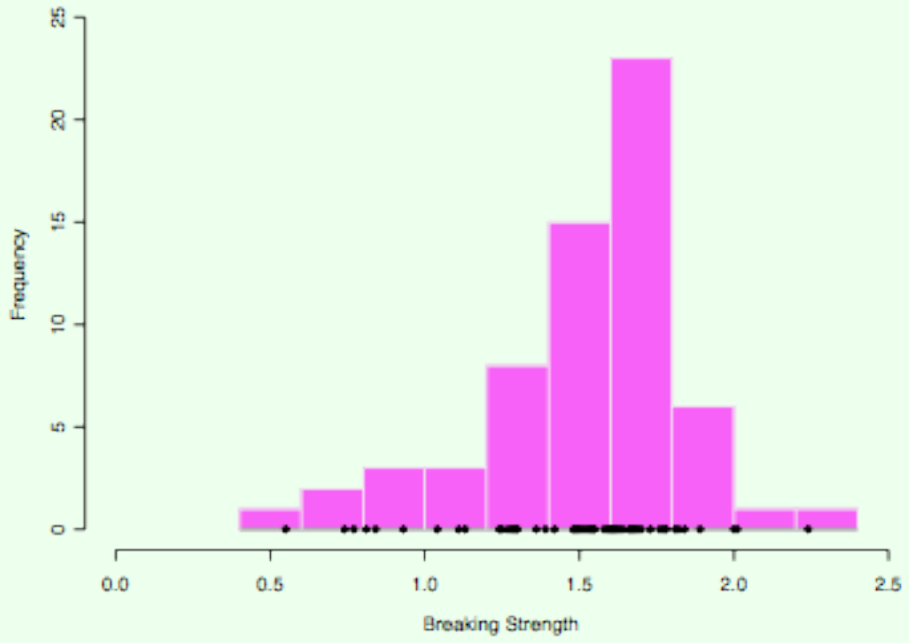
Para estas situações, as inferências usuais baseadas no comportamento médio de populações aproximadamente normais não são adequadas;

Uma forma específica de análise, denominada *análise de valores extremos*, se faz necessária para lidar com as características específicas deste tipo de dados:

- Baixa disponibilidade de observações: valores extremos são mais raramente observados que valores próximos à média, *para a maior parte das distribuições usuais de probabilidade*;
- Distribuições específicas: a distribuição dos valores extremos usualmente não segue a distribuição geral dos dados;

Exemplo

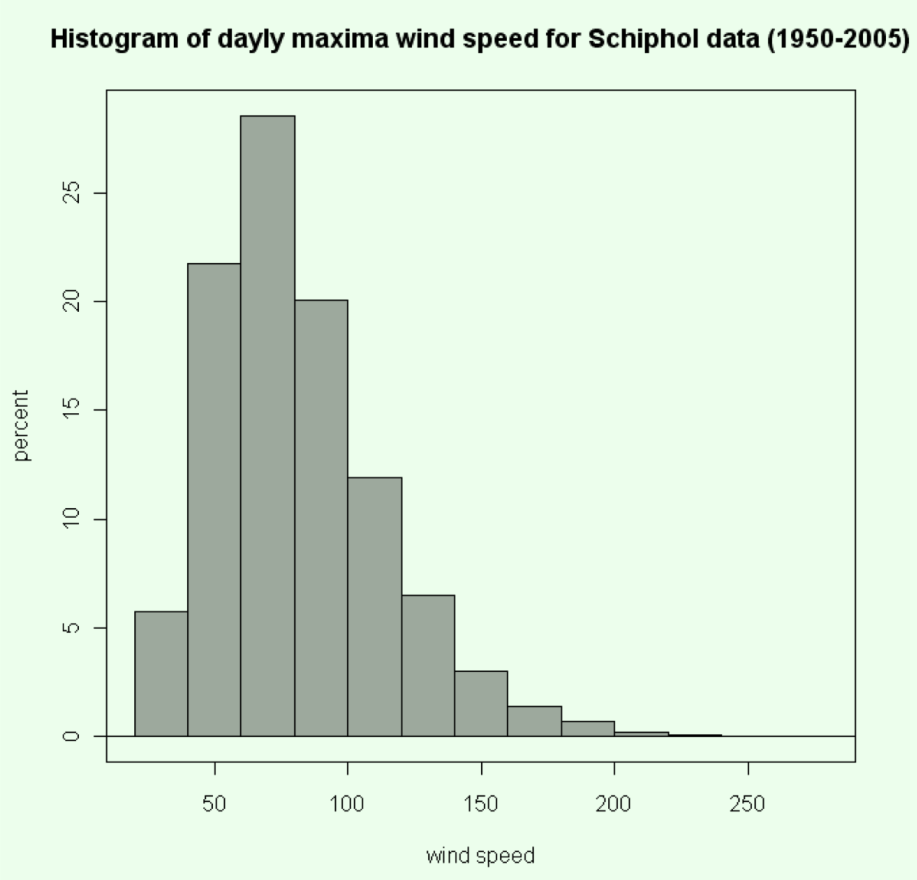
Força de ruptura em fibras de vidro.



(adaptado de Smith&Naylor, 1987)

Exemplo

Velocidade máxima de ventos.



(adaptado de Omey, Mallor&Nualart, 2009)

Modelagem de extremos por blocos

Conceitos iniciais

Historicamente, a área de modelagem de valores extremos teve início com o tratamento de valores máximos divididos em blocos:

- Máximos anuais de eventos meteorológicos;
- Desempenhos mínimos de lotes de componentes;

Embora hajam atualmente modelos capazes de uma melhor utilização de dados para a modelagem de extremos, é interessante compreender os conceitos e modelos utilizados para esta modelagem de forma a adquirir uma base mais sólida para a compreensão das demais técnicas.

Modelagem de extremos por blocos

Formulação do modelo

Os modelos utilizados para descrever os valores máximos[1] de sequências aleatórias tem como foco o comportamento estatístico de variáveis do tipo:

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

onde os X_i são variáveis aleatórias identica e independentemente distribuídas.

Para que seja possível tratar estes valores utilizando modelos assintóticos, é necessária sua padronização de acordo com:

$$M_n^* = \frac{M_n - b_n}{a_n}$$

para constantes reais finitas $a_n > 0$ e b_n.

[1] Modelos de mínimos podem ser desenvolvidos de forma análoga.

Modelagem de extremos por blocos

Teorema dos Tipos Extremos

O teorema dos tipos extremos é um resultado estatístico extremamente útil, considerado como sendo o análogo do *teorema do limite central* para dados extremos. Este teorema dita que, se existirem constantes $a_n > 0$ e b_n de forma que:

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

onde G é uma distribuição não-degenerada de probabilidades, então a distribuição de G pertence a uma de três famílias de distribuições, *independentemente da distribuição original dos dados*. Estas famílias são conhecidas como *distribuições de vaores extremos* (EVDs).

Modelagem de extremos por blocos

Tipo Extremo I - distribuição de Gumbel

Distribuições de valores extremos do tipo I são modelos para as caudas inferior e superior de distribuições do tipo exponencial:

- Gaussiana;
- Lognormal;
- Exponencial;
- etc.

A p.d.f. para esta distribuição é dada por:

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

Modelagem de extremos por blocos

Tipo Extremo II - distribuição de Fréchet

Distribuições de valores extremos do tipo II não possuem grande importância no contexto de engenharia de confiabilidade.A distribuição de Fréchet é descrita por:

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

Modelagem de extremos por blocos

Tipo Extremo III - distribuição de Weibull

Distribuições de valores extremos do tipo II são modelos para os mínimos de distribuições que são inferiormente limitadas. Estes valores são descritos pela distribuição de Weibull,

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

Uma das mais utilizadas em engenharia de confiabilidade, para a modelagem de funções de confiabilidade, distribuição de resistência de materiais, etc.

Modelagem de extremos por blocos

Distribuições de valores extremos

Cada um dos três tipos de distribuição descritas pelo teorema dos tipos extremos possui um parâmetro de localização *a* e um parâmetro de escala *b*. Além disso, as distribuições de Fréchet e Weibull possuem um parâmetro adicional de forma, α.

Nos primeiros estudos de valores extremos, era usual assumir uma das três famílias de distribuições, estimar seus parâmetros e realizar a análise a partir desta escolha inicial.

Entretanto, tal estratégia implicava em uma fonte adicional de erros e de vieses na análise.

Modelagem de extremos por blocos

Distribuição generalizada de valores extremos

Uma análise menos sujeita a estes problemas pode ser obtida a partir de uma reformulação do teorema dos tipos extremos, na qual as três distribuições de valores extremos são interpretadas como casos particulares de uma única distribuição, conhecida como *distribuição generalizada de valores extremos*:

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

definida nos valores $\{z \in \mathbb{R} | 1+\xi(z-\mu)/\sigma > 0\}$, e com parâmetros satisfazendo as condições:

- $-\infty < \mu < \infty$ - parâmetro de localização;
- $\sigma > 0$ - parâmetro de escala;
- $-\infty < \xi < \infty$ - parâmetro de forma.

Modelagem de extremos por blocos

Distribuição generalizada de valores extremos

Esta unificação das três distribuições de valores extremos em uma única família simplifica grandemente o processo de definição de modelos adequados, uma vez que a responsabilidade pela escolha do modelo repousa nos *dados*, ao invés de no analista.

O parâmetro de forma ξ essencialmente determina o tipo da distribuição:

- $\xi < 0$ - Weibull
- $\xi > 0$ - Fréchet
- $\xi \to 0$ - Gumbel

Além disso, a incerteza associada à definição deste parâmetro quantifica a falta de dados suficientes para a escolha de uma das três EVDs.

Modelagem de extremos por blocos

Distribuição generalizada de valores extremos - valores mínimos

No caso da necessidade de modelagem de valores mínimos, a distribuição GEV pode ser descrita como:

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

definida nos valores $\{z \in \mathbb{R} | 1 + \xi(z - \tilde{\mu})/\sigma > 0\}$, com:

- $-\infty < \tilde{\mu} < \infty$ - parâmetro de localização;
- $\sigma > 0$ - parâmetro de escala;
- $-\infty < \xi < \infty$ - parâmetro de forma.

Alternativamente, pode-se utilizar a transformação de dados $\tilde{M}_n = -\max(-X_1, \ldots, -X_n)$ e aplicar diretamente o modelo de máximos.

Modelagem de extremos por blocos

Estimação dos parâmetros da distribuição

A estimação dos parâmetros $\mu,\ \sigma,\ \xi$ da distribuição generalizada de valores extremos pode ser realizada de diversas formas (assim como a estimativa dos valores de médias e desvios-padrão das distribuições usuais);

Uma das formas mais usuais de realizar esta estimação é a utilização do método de máxima verossimilhança (*maximum likelihood*, ML).

Este método é baseado na maximização da função *likelihood* (u de seu logaritmo), que quantifica a probabilidade de se obter os dados observados caso o modelo utilizado seja verdadeiro para dados valores dos parâmetros.

Modelagem de extremos por blocos

Estimação dos parâmetros da distribuição

Supondo que $\mathbf{Z} = [Z_1, \ldots, Z_m]$ é um vetor de variáveis aleatórias i.i.d. seguindo a GEV, a função *log-likelihood* para os parâmetros da GEV com $\xi \neq 0$ é dada por:

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

sob a condição de que:

$$1 + \xi \left(\frac{z_i - \mu}{\sigma} \right) > 0, \ \forall i \in \{1, \ldots, m\}$$

Modelagem de extremos por blocos

Estimação dos parâmetros da distribuição

Para o caso onde $\xi = 0$, a função é obtida a partir da distribuição limite de Gumbel:

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

A maximização destas equações em relação ao vetor de parâmetros (μ, σ, ξ) conduz às estimativas de máxima verossimilhança dos parâmetros do modelo.

Modelagem de extremos por blocos

Estimação dos parâmetros da distribuição

Embora não haja solução analítica para esta maximização, a mesma é facilmente realizada a partir de algoritmos usuais de otimização não-linear restrita, haja vista a necessidade de obter valores que obedeçam a condição

$$1 + \xi \left(\frac{z_i - \mu}{\sigma} \right) > 0, \ \forall i \in \{1, \ldots, m\}$$

e que evitem a avaliação da primeira equação na vizinhança de pontos onde $\xi \approx 0$. Esta última condição é facilmente atendida através da modificação da função avaliada para uma dada ϵ-vizinhança de $\xi = 0$.

Modelagem de extremos por blocos

Inferência utilizando a GEV

Uma vez que valores apropriados para os parâmetros da GEV tenham sido obtidos, a mesma pode ser utilizada para a geração de predições a respeito dos valores extremos.

Um conceito importante neste tipo de inferência é o *nível de retorno*:

Um *nível de retorno* z_p, associado a um *período de retorno* $1/p$, $p \in [0, 1]$, representa um valor da variável observada que se espera que seja excedido uma vez a cada $1/p$ observações.

Neste contexto, cada observação representa o valor máximo de blocos do tamanho originalmente utilizado. Por exemplo, em inferências sobre máximos meteorológicos anuais, $1/p$ é um valor dado em anos.

Modelagem de extremos por blocos

Inferência utilizando a GEV

Dadas as estimativas de máxima verossimilhança $\left(\hat{\mu}, \hat{\sigma}, \hat{\xi}\right)$, a estimativa correspondente para um dado nível de retorno z_p pode ser facilmente calculada por:

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

com $y_p = -\log(1-p)$. Intervalos de confiança nesta predição também podem ser obtidos de forma relativamente simples utilizando um médoto conhecido como *método delta*.

Modelagem de extremos por blocos

Solução computacional

Embora seja relativamente simples derivar os valores dos parâmetros e os níveis de retorno analiticamente, soluções computacionais são frequentemente mais simples e mais diretas.

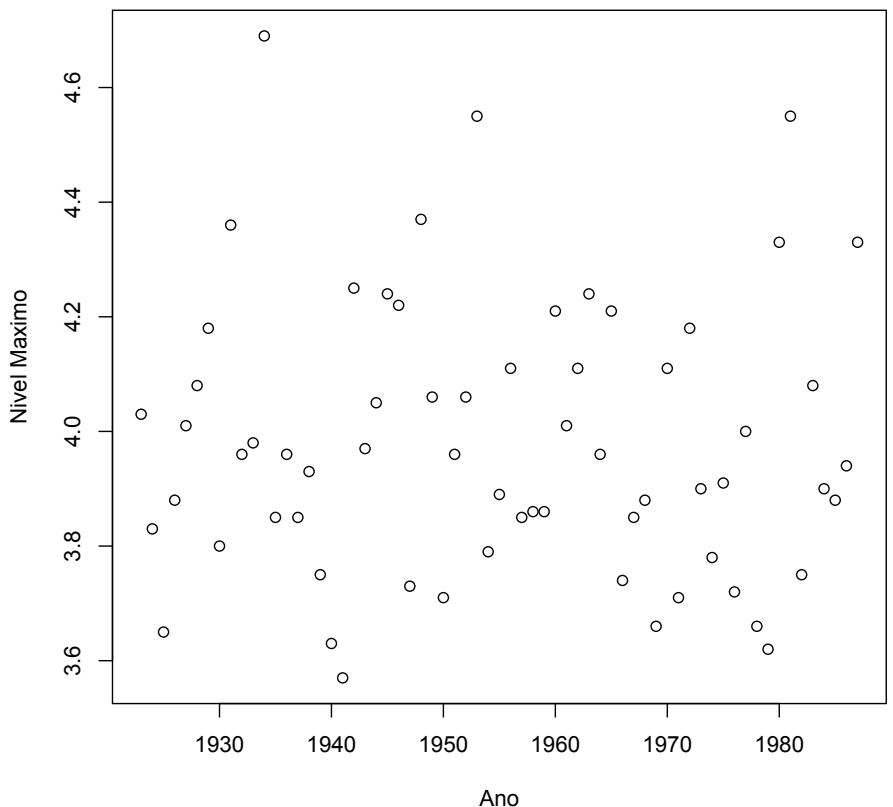
Diversos pacotes do software estatístico **R** possuem rotinas para a inferência sobre valores extremos, como por exemplo:

- *ismev*;
- *evir*;
- *extRemes*;
- *evd*;
- etc.

Modelagem de extremos por blocos

Solução computacional

Como um exemplo de aplicação, suponhamos que para a especificação de um sistema de escoamento de dejetos líquidos em uma cidade litorânea seja necessário estimar o nível máximo do mar nos próximos 50 anos. suponha ainda que tenhamos dados dos máximos anuais dos últimos 64 anos.



Modelagem de extremos por blocos

Solução computacional

```
# Carrega biblioteca
library(ismev)

# Carrega dados
data(portpirie)
summary(portpirie)
```

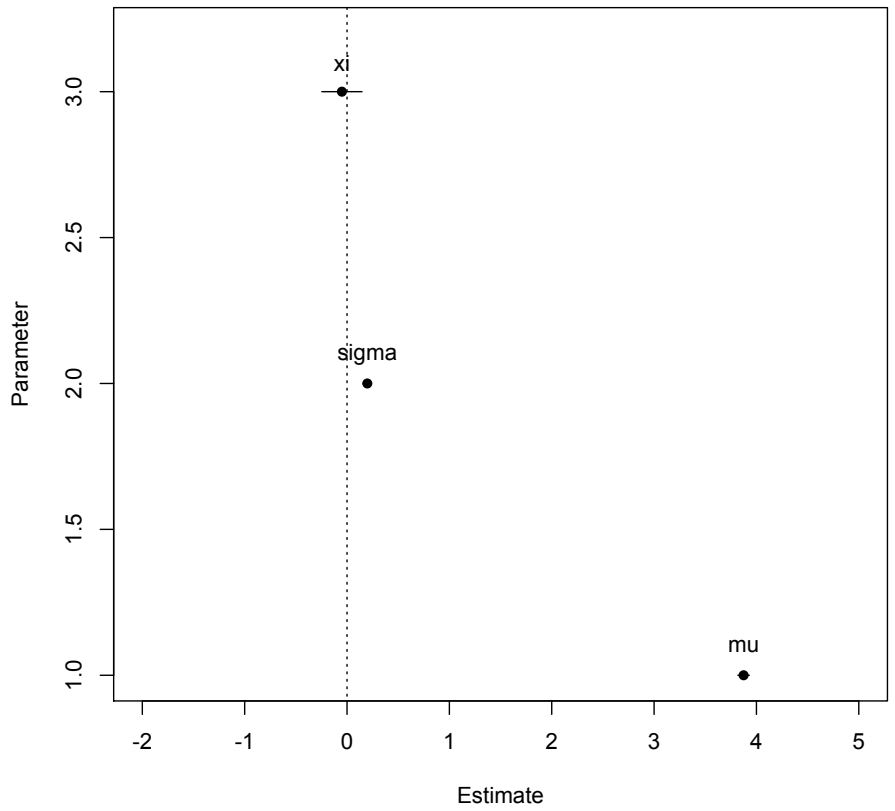
```
      Year          SeaLevel
 Min.   :1923   Min.   :3.570
 1st Qu.:1939   1st Qu.:3.830
 Median :1955   Median :3.960
 Mean   :1955   Mean   :3.981
 3rd Qu.:1971   3rd Qu.:4.110
 Max.   :1987   Max.   :4.690
```

```
 # Ajusta modelo GEV
 model<-gev.fit(portpirie[,2])
```

Modelagem de extremos por blocos

Solução computacional

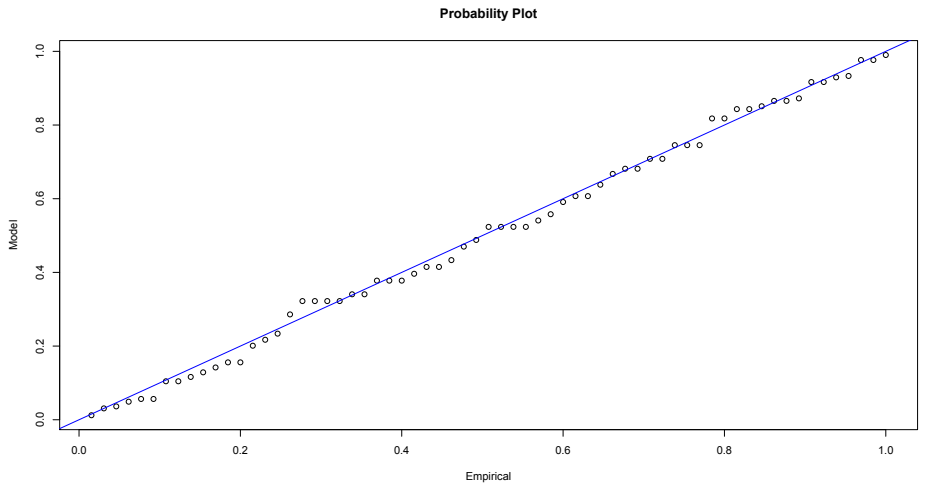
```
# Plota valores dos parâmetros estimados
est<-model$mle
se<-model$se

# Plot estimates
plot(est,1:3,type="p",xlim=c(-2,5),
ylim=c(1,3.2),xlab="Estimate",
ylab="Parameter",pch=16)
for (i in 1:3){
points(x=c(est[i]-2*se[i],
est[i]+2*se[i]),
y=c(i,i),type="l")
}
points(c(0,0),c(0,4),type="l",lty=3)
text(est,0.1+(1:3),
labels=c("mu","sigma","xi"))
```

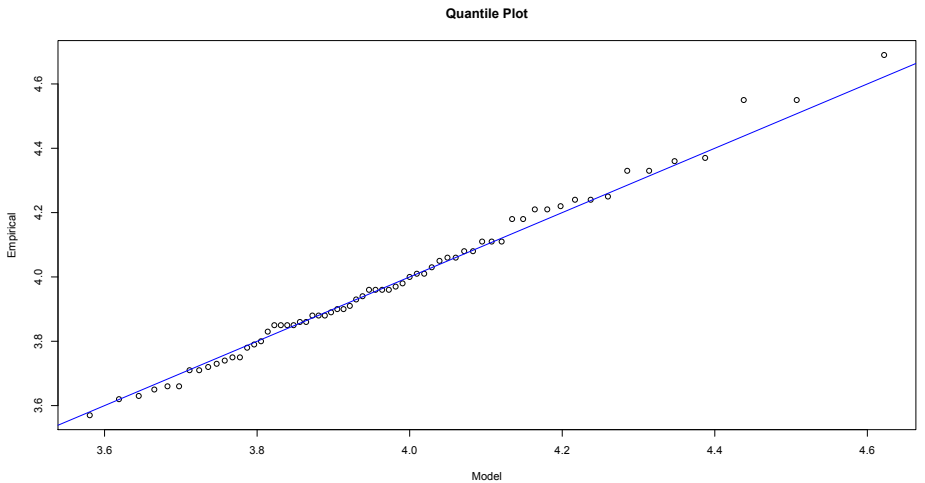


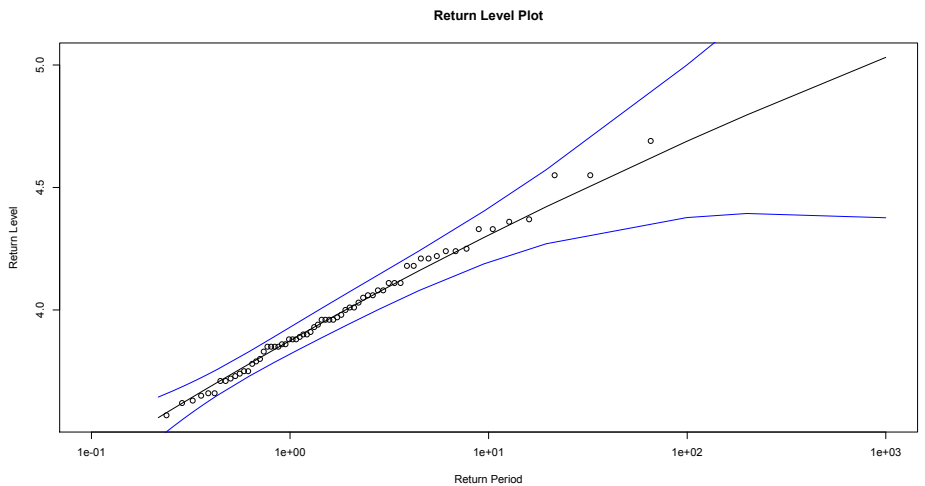
Modelagem de extremos por blocos

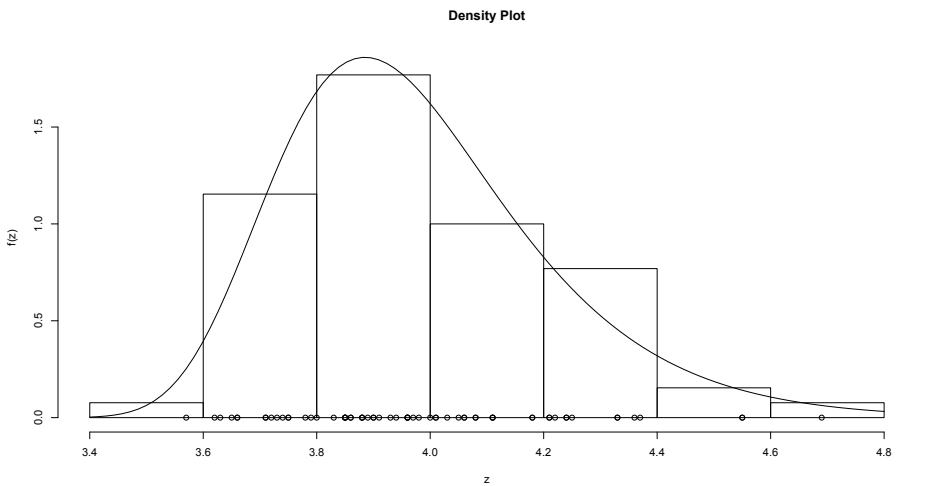
Solução computacional

```
# Plota diagnósticos do modelo
gev.diag(model)
```





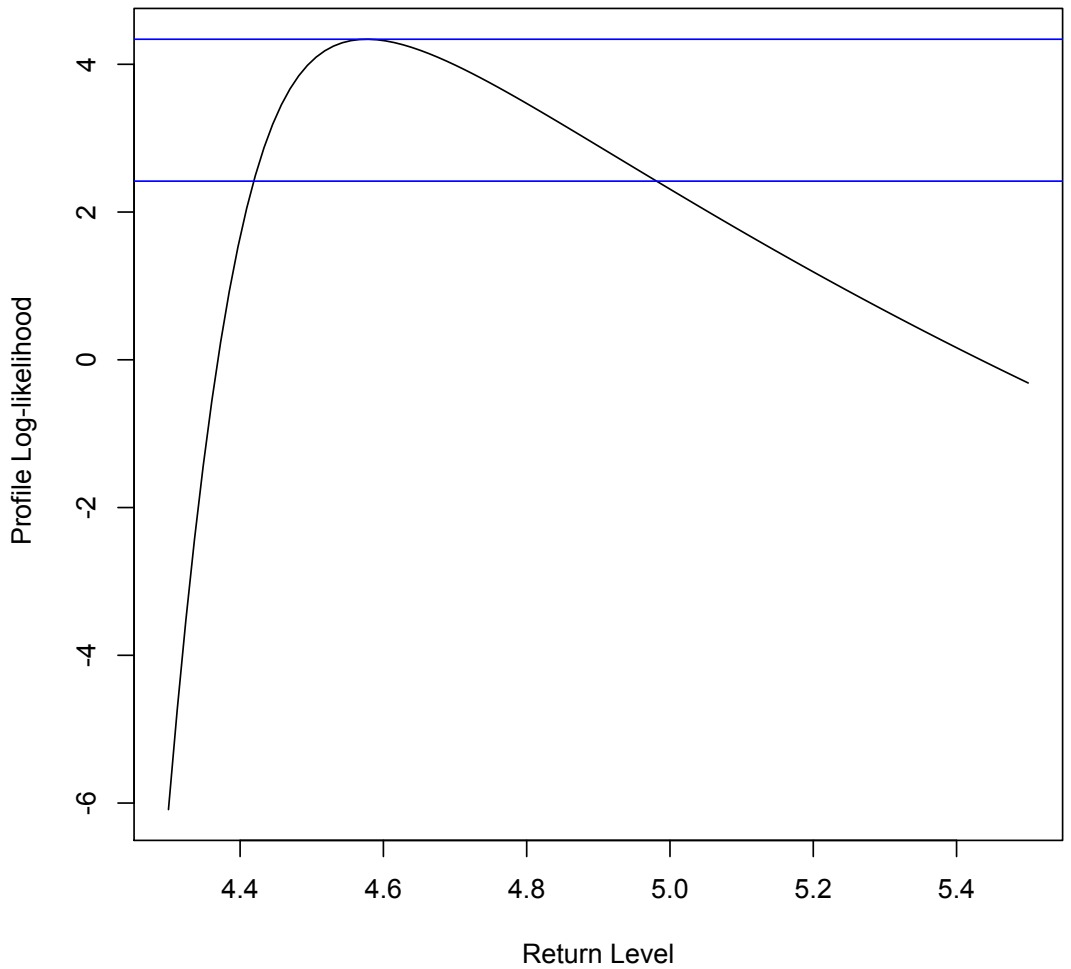




Modelagem de extremos por blocos

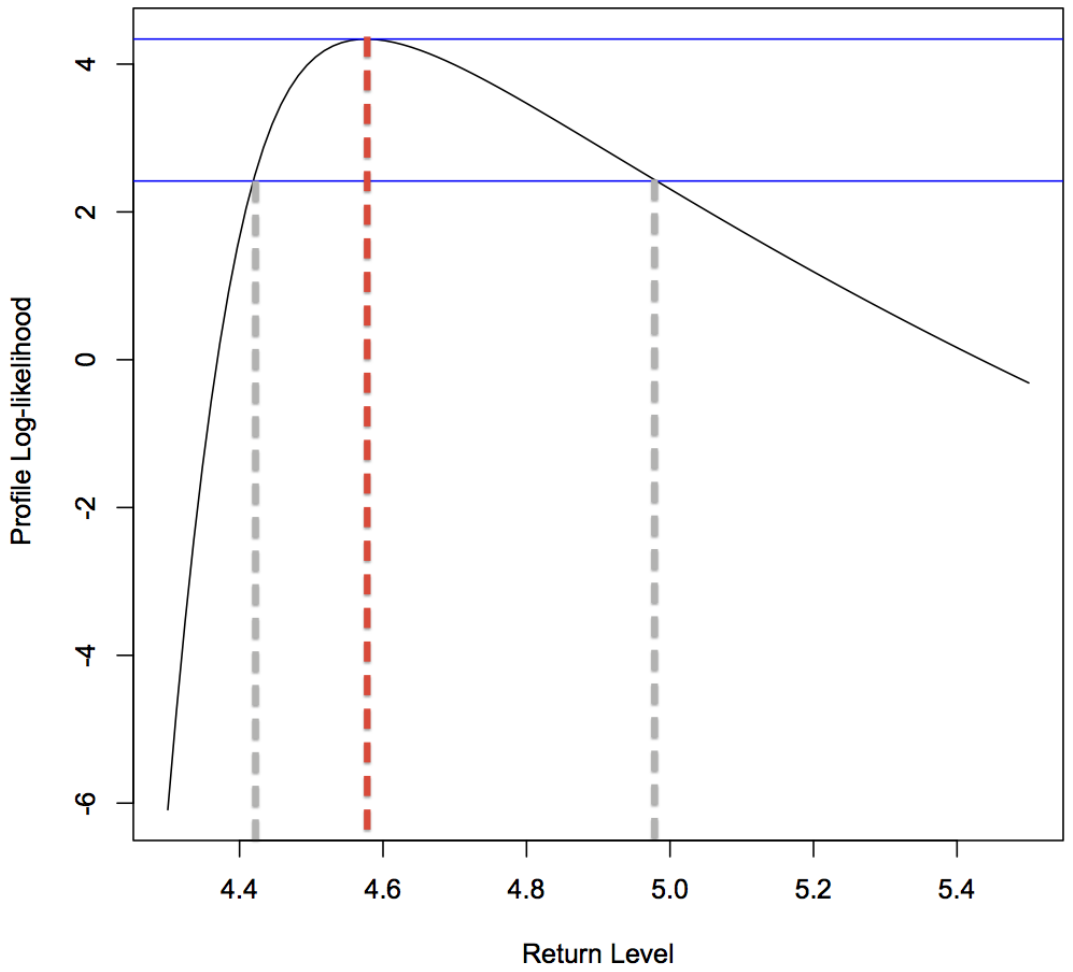
Solução computacional

```
# Nível de retorno para 50 anos
gev.prof(model,50,xlow=4.3,xup=5.5)
```



Modelagem de extremos por blocos

Solução computacional



Modelos de Limiar

Conceitos iniciais

Embora útil, a modelagem de extremos utilizando pontos máximos de blocos de observações é uma abordagem que leva ao desperdício de informações, principalmente em casos onde outros dados relativos a comportamentos extremos estão disponíveis.

Por exemplo, se um bloco contém muitas observações mais extremas que outros, a informação sobre este comportamento é perdida.

Se uma sequência de valores está disponível, um uso mais eficaz dos dados pode ser obtida evitando-se o procedimento de agrupamento em blocos.

Modelos de Limiar

Conceitos iniciais

Seja $\mathbf{X} = [X_1, X_2, \ldots]$ uma sequência de variáveis aleatórias i.i.d. de acordo com uma dada distribuição F. Uma interpretação natural de observações extremas é considerá-los como *valores que excedem um determinado limiar elevado u*.

De acordo com esta definição, o comportamento de observações extremas pode ser descrito por:

$$P\{X_i > u + y | X_i > u\} = \frac{1 - F(u + y)}{1 - F(u)}, \quad y > 0$$

Se a distribuição F for conhecida, a distribuição de valores acima do limiar também será.

Modelos de Limiar

Distribuição generalizada de Pareto

Um outro resultado, entretanto, dispensa o conhecimento da distribuição exata dos X_i. Pode-se provar que, para a sequência $\mathbf{X}$ descrita anteriormente, a distribuição de valores $(X_i - u)$ condicional a $X > u$ pode ser aproximada por:

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

para $y > 0$ e sujeita a $(1 + \xi y/\tilde{\sigma}) > 0$, com:

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

Esta família de distribuições é conhecida como *família de Pareto generalizada*.

Modelos de Limiar

Distribuição generalizada de Pareto

Assim como no caso da GEV, o parâmetro de forma ξ possui um papel central na determinação do comportamento da GPD:

- $\xi < 0$ - a distribuição de excessos possui um limite superior dado por $u - \tilde{\sigma}/\xi$;
- $\xi > 0$ - a distribuição nao possui limite superior;
- $\xi \to 0$ - A distribuição se torna uma distribuição exponencial,

$$H(y) = 1 - \exp\left(-\frac{y}{\tilde{\sigma}}\right), \quad y > 0$$

Modelos de Limiar

Escolha do limiar

A escolha do valor de limiar é um ponto crítico para a boa modelagem de valores extremos:

- Valores baixos de u tendem a violar as propriedades assintóticas que formam a base do modelo, levando à ocorrência de erros sistemáticos no modelo (bias);
- Valores muito elevados de u resultam na ocorrência de poucos exemplos de pontos classificados como extremos, levando a um aumento na incerteza dos parâmetros (elevação na variância).

O procedimento padrão envolve tentar escolher um valor de u o mais baixo possível, sujeito à condição de que o modelo resultante seja capaz de gerar boas aproximações dos dados.

Modelos de Limiar

Escolha do limiar

Um método frequentemente utilizado para a seleção do valor de u é baseado no valor da média da distribuição generalizada de Pareto.

Supondo uma variável aleatória Y que siga uma GPD com parâmetros σ e ξ, sabe-se que:

$$E[Y] = \frac{\sigma}{1 - \xi}$$

para $\xi < 1$. Valores de $\xi \geq 1$ levam a valores de média infinitos.

Modelos de Limiar

Escolha do limiar

Assumindo que a GPD seja um modelo adequado para os valores em excesso de um dado limiar u_0 gerados por uma série $X_1, \ldots, X_n$. De acordo com a equação anterior, temos que:

$$E\left[X_i - u_0 | X_i > u_0\right] = \frac{\sigma_{u_0}}{1-\xi}, \quad \text{para } \xi < 1.$$

Pode-se mostrar que se a GPD é um modelo adequado para um limiar u_0, também será para qualquer limiar $u > u_0$, sujeito a mudanças no fator de escala σ_u e que, para $u > u_0$:

$$E\left[X_i - u | X_i > u\right] = \frac{\sigma_u}{1-\xi} = \frac{\sigma_{u_0} + \xi u}{1-\xi}$$

representa a média dos valores de excesso em relação a u.

De acordo com este último resultado, tem-se que o valor da média dos valores de excesso em relação a u tende a crescer linearmente com o aumento no valor de u (para faixas de valores nas quais a GPD seja apropriada). Isto sugere um procedimento baseado no gráfico de valores de u versus:

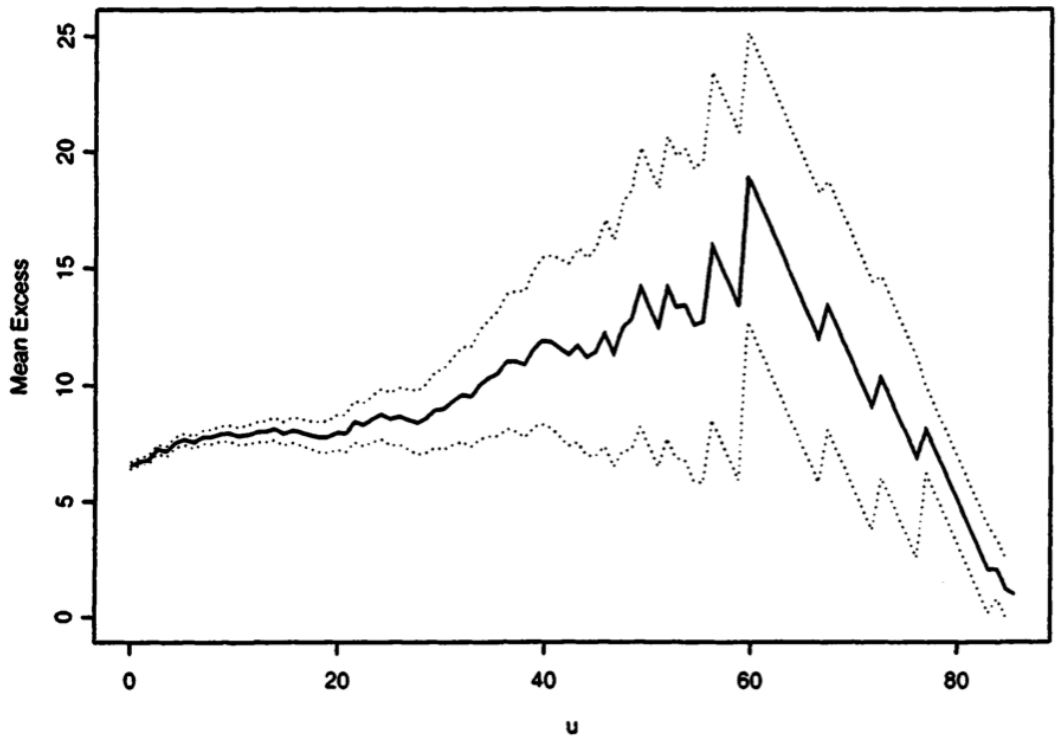
$$\frac{1}{n_u} \sum_{i=1}^{n_u} \left(x_{(i)} - u\right)$$

onde $x_{(i)}$ representamas n_u observações que excedem o limiar u, x_{max} é o maior valor observado, e $u < x_{max}$. Este gráfico é conhecido como *gráfico de vida média residual* (*mean residual life plot*).

Modelos de Limiar

Escolha do limiar

Este gráfico é conhecido como *gráfico de vida média residual* (*mean residual life plot*). Acima de um dado limiar u_0 para o qual a GPD passa a ser um modelo válido, este plot deve ter características aproximadamente lineares (com inclinação negativa) em relação a u.



Modelos de Limiar

Estimação dos parâmetros do modelo

Os parâmetros do modelo GPD podem ser estimados da mesma forma que os da GEV, ou seja, através da maximização da função log-likelihood. Para $\xi \neq 0$:

$$\ell(\sigma, \xi) = k \log \sigma - (1 + 1/\xi) \sum_{i=1}^{k} \log (1 + \xi y_i / \sigma)$$

para $(1 + \xi y_i/\sigma) > 0 \; \forall i \in (1, \ldots, k)$. Para $\xi = 0$,

$$\ell(\sigma) = -k \log \sigma - \frac{1}{\sigma} \sum_{i=1}^{k} y_i$$

Novamente, métodos numéricos de otimização são utilizados para encontrar o valor do máximo destas funções.

Modelos de Limiar

Níveis de Retorno

Assim como nos modelos GEV, os níveis de retorno x_m quantificam valores que se espera serem excedidos a cada m observações. Pode-se mostrar que estes níveis de retorno são calculáveis a partir de:

$$x_m = \begin{cases} u + \frac{\sigma}{\xi}\left[\left(\frac{mk}{n}\right)^{\xi} - 1\right] & , \ \xi \neq 0 \\ u + \sigma \log\left(\frac{mk}{n}\right) & , \ \xi = 0 \end{cases}$$

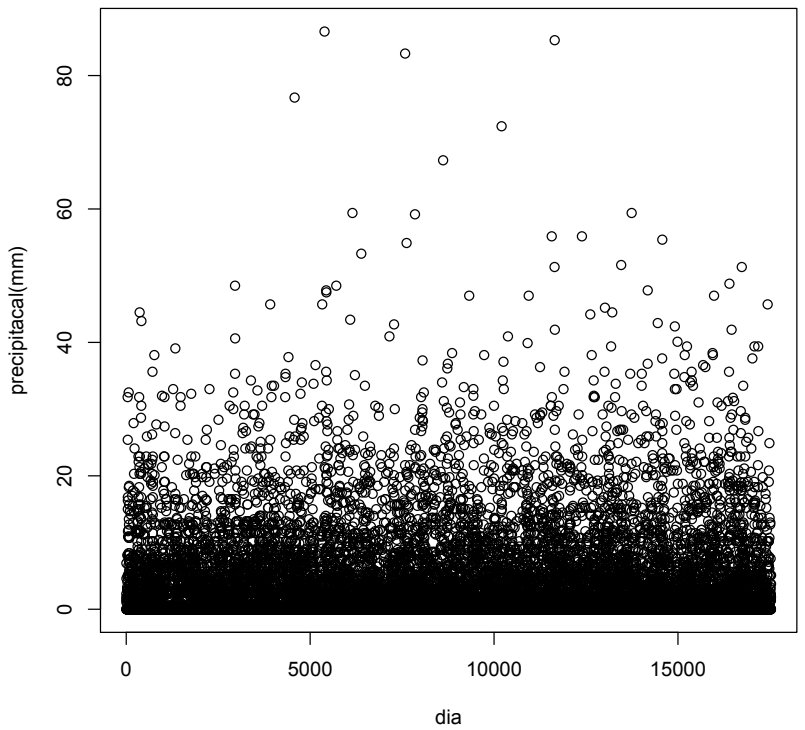
com k/n representando a proporção de observações acima do limiar u, e assumindo $x_m > u$.

Por construção, o valor de x_m pode ser considerado como sendo o *nível de retorno após m observações*. Intervalos de confiança nestes valores também podem ser facilmente obtidos em software.

Modelagem de extremos por blocos

Solução computacional

Suponhamos agora que seja necessário projetar um sistema de escoamento pluvial capaz de suportar o máximo de precipitação diária previsível para uma dada região da cidade. Posto que obras deste porte são custosas e raras, deseja-se projetar o sistema de forma a possibilitar o escoamento da maior precipitação esperada para os próximos 100 anos.



Modelagem de extremos por blocos

Solução computacional

Selecionando arbitrariamente $u = 30$, podemos ajustar o modelo:

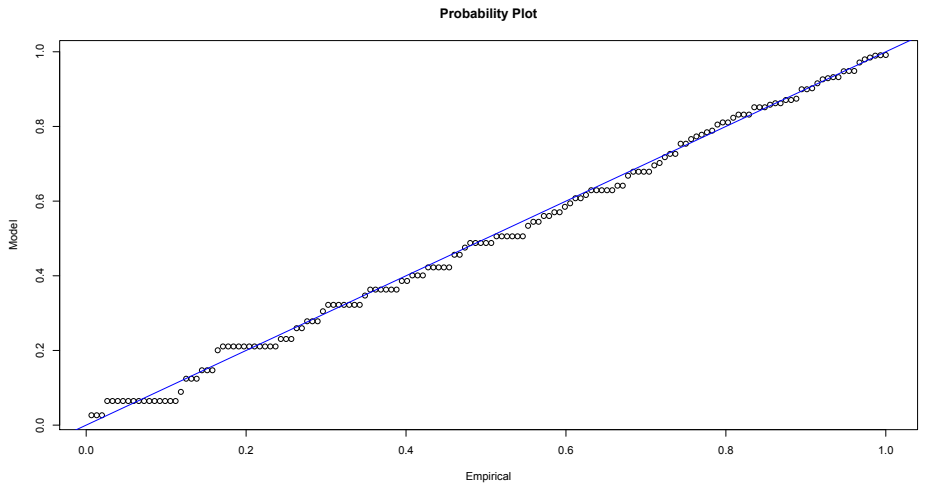
```
model<-gpd.fit(rain,30)

$nexc
[1] 152

$mle
[1] 7.4422639 0.1843027

$se
[1] 0.9587773 0.1011714
```

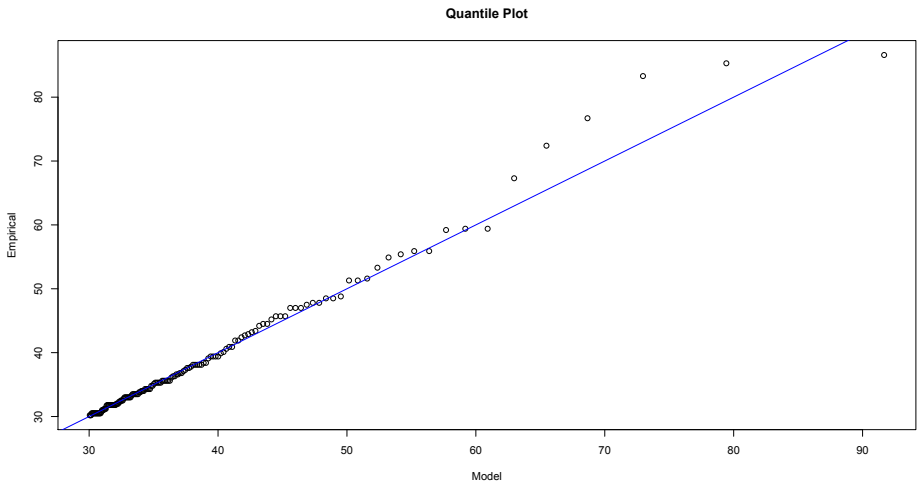
Modelagem de extremos por blocos

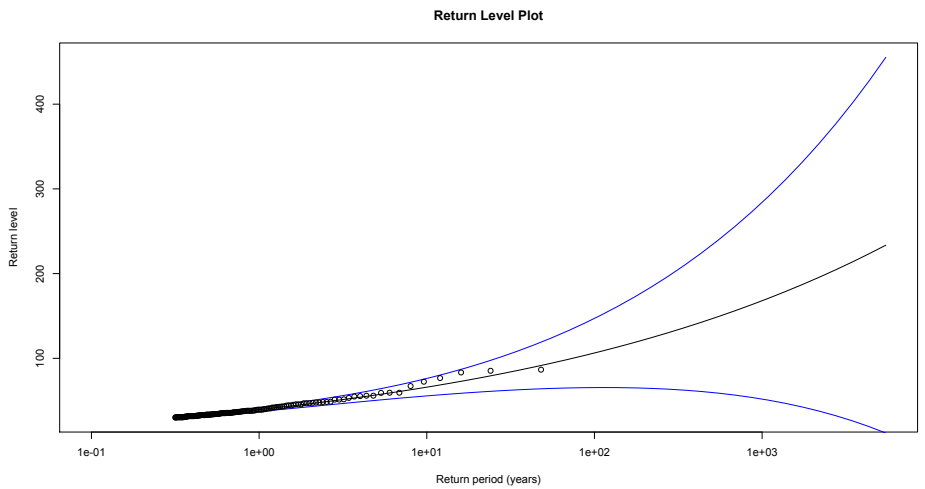
Solução computacional

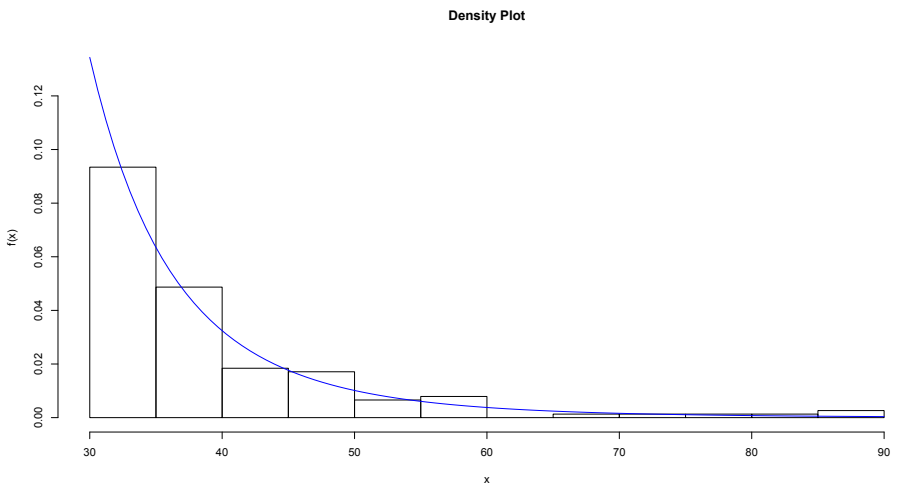
Gerando os gráficos de diagnóstico, temos:

```
gpd.diag(model)
```







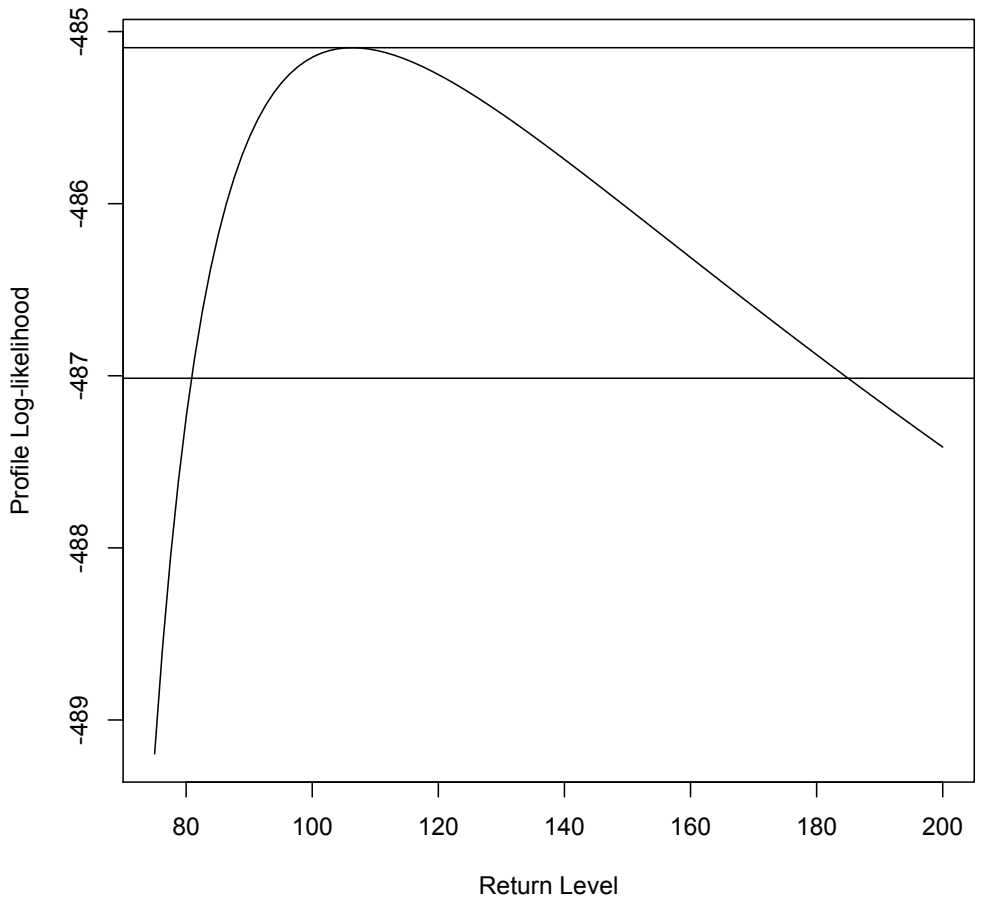


Modelagem de extremos por blocos

Solução computacional

Finalmente, o nível de retorno para 100 anos

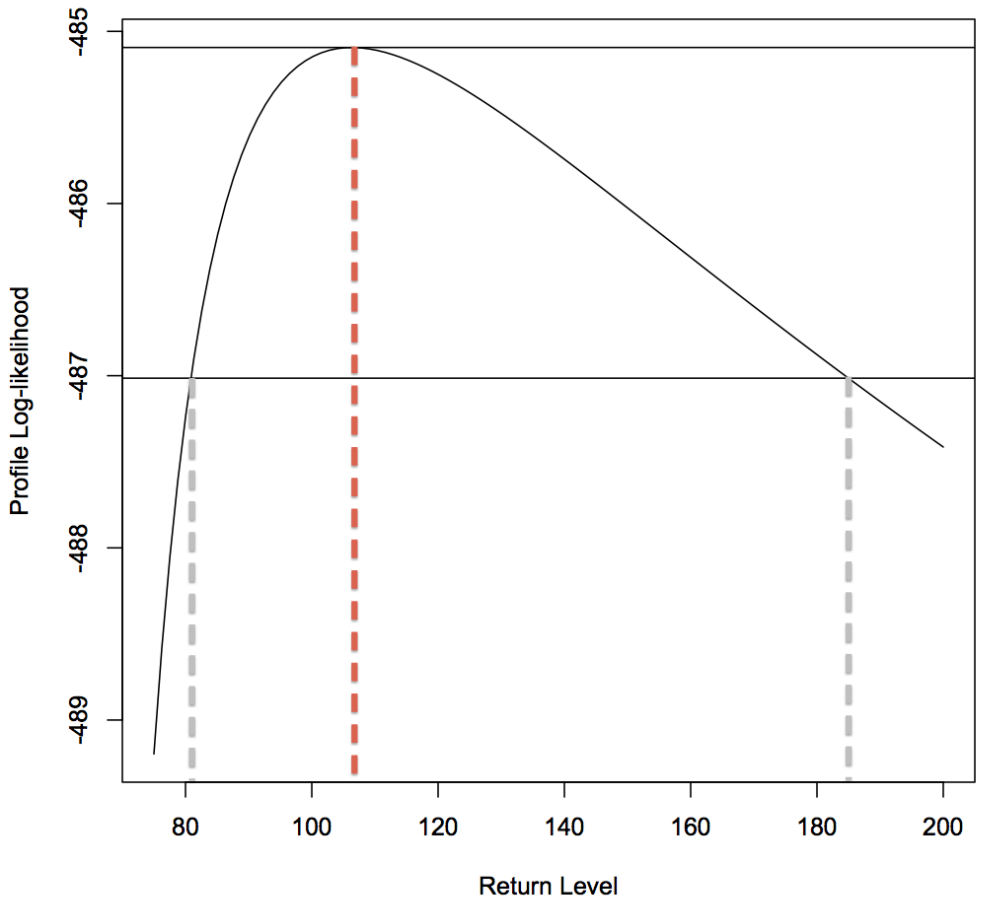
```
gpd.prof(model)
```



Modelagem de extremos por blocos

Solução computacional

Finalmente, o nível de retorno para 100 anos



Conclusões

A modelagem de valores extremos, embora aparentemente complexa, pode ser realizada de forma relativamente direta através da utilização de ferramentas computacionais

Este tipo de modelagem é extremamente útil para a estimativa de comportamentos de pior caso - sejam mínimos ou máximos - o que representa uma tarefa essencial na derivação de conceitos mais avançados de confiabilidade.

Desafio: realizar a modelagem dos valores mínimos de resistência mecânica de fibras de vidro contidas no *dataset* **glass** do pacote **ismev**.

Bibliografia

Referências utilizadas

1. P.D.T. O'Connor, A. Kleyner, *Practical Reliability Engineering*, 5th ed., Wiley, 2012 - cap. 2;
2. S. Coles, *An introduction to the statistical modeling of extreme values*, 1st ed., Springer 2001 - Caps. 1 - 4;
3. S. Coles, A. Davison, *Statistical Modeling of Extreme Values*, 2008 - `http://goo.gl/cLMg3C`
4. Pacote ismev - `http://CRAN.R-project.org/package=ismev`