2^a Prova - FECD B - 2025/01

Renato Assunção

Problem: Modeling File Download Sizes with a Pareto Distribution

Background

A Pareto probability distribution is used when we have a heavy-tail phenomenon: positive random values such that most of them are small but some can be large and it is not rare to see very, very large values. Think of income distribution or the distribution of number of followers in a large social network. In these analysis, we focus on the large values and hence we ignore the vary small values by looking only at values above a certain threshold x_m. A Pareto distribution for the continuous variable Y has density function given by:

$$f(y;\alpha,x_m) = \begin{cases} \frac{\alpha x_m^{\alpha}}{y^{\alpha+1}}, & \text{if } y > x_m \\ \\ 0, & \text{if } y \leq x_m \end{cases}$$

The left-hand side of Figure 1 shows the typical shape of this density function (the plot uses $\alpha = 2.3$ and $x_m = 100$). The center plot is a histogram of a sample with 100 values from this density. The right-hand side plot shows $\mathbb{E}(Y)$ as a function of α (for $x_m = 100$).

In contrast with the most common distributions (Gaussian, Gamma, Weibull, etc.), which have an exponential decay as y becomes sufficiently large, the Pareto distribution has a polynomial decay as y increases. For example, if $x_m = 1$ and $\alpha = 2$, the density over the support set ($y > x_m = 1$) is given by $f(y) = 2/y^3$, a decay according to a polynomial of third degree. This means that, rather than a fast (exponential) decrease in the probability of large values, the Pareto distribution decays slowly (in a polynomial pace), and hence large (and very large) values are still relatively common in a sample. The notion of a "very large value" refers to values that are orders of magnitude greater than the expected value $\mathbb{E}(Y)$.

The scale or threshold parameter x_m must be positive and it is the minimum possible value of the random variable Y. In this problem, x_m is known and fixed at $x_m = 100$.

The shape parameter α must also be positive. It governs the heaviness of the tail of the distribution. Small α leads to a distribution where extremely large values appear easily (heavy tail distributions). This parameter controls whether the mean or variance exists: $\mathbb{E}(Y)$ exists only if $\alpha > 1$; $\mathbb{V}(Y)$ exists only if $\alpha > 2$.

Context

A cloud storage company wants to model the file sizes (in MB) downloaded by different users in a given month. These file sizes are believed to follow a heavy-tailed distribution (many small files and some large or very, very large files). We monitored only the download of files larger than 100 MB (that is, $x_m = 100$). Each heavy-user i has a download size Y_i modeled as a Pareto random variable with probability density function with unknown shape parameter α_i:

$$f(y;\alpha_i,x_m = 100) = \begin{cases} \frac{\alpha 100^{\alpha_i}}{y^{\alpha_i+1}}, & \text{if } y > 100 \\ \\ 0, & \text{if } y \leq 100 \end{cases}$$

In our problem, the shape parameter α_i varies among the users and it depends on features or covariates via:

$$\alpha_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}) = \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})$$

where

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \quad \boldsymbol{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \end{pmatrix}$$

are column-vectors of dimension 3×1 with the features in $\boldsymbol{x}_i$ given by:

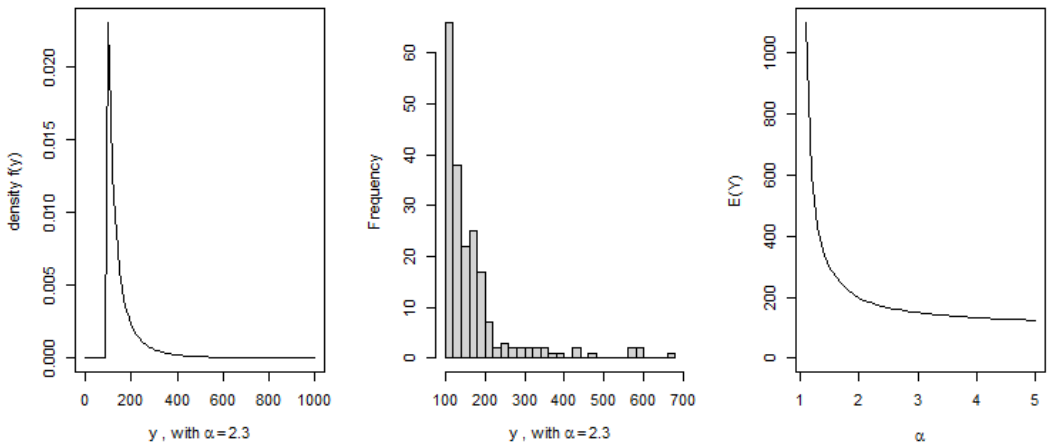

Figura 1: Left: Pareto density with $\alpha = 2.3$. Center: Histogram of a sample of 200 values from the Pareto with $\alpha = 2.3$. Right: Plot of $\mathbb{E}(Y)$ as a function of α.

- $x_{i1} = 1$ if the user i is a developer, and $= 0$, otherwise
- x_{i2} is the number of devices used to access the service by user i

The expected value exists (when $\alpha_i > 1$) and is given by:

$$\mathbb{E}(Y_i) = \frac{x_m \alpha_i}{\alpha_i - 1} = \frac{100\alpha_i}{\alpha_i - 1}$$

Sample Data for $n = 10$ users

Variable	1	2	3	4	5	6	7	8	9	10
y_i (MB)	180	150	560	220	130	300	170	155	390	145
x_{i1}	1	0	1	1	0	0	1	0	1	0
x_{i2}	3	2	5	2	1	4	3	2	4	2

Questions

(a) Assume that these data have been generated with the true value of the parameter equal to $\boldsymbol{\beta} = (0.3, -0.5, 0.1)$. Write the mathematical expressions for $\mathbb{E}(Y_i)$ as a function of x_2 for both developer and non-developer users (that is, the expressions for $\mathbb{E}(Y_i)$ as a function of x_2 when $x_1 = 0$ and $x_1 = 1$).

(b) Derive the log-likelihood function.

(c) Compute the score vector $\nabla \ell(\boldsymbol{\beta})$.

(d) Compute the element $(2, 2)$ of the Hessian matrix $\nabla^2 \ell(\boldsymbol{\beta})$.

(e) When Y is a Pareto distribution with parameter α and threshold $x_m = 100$, it is possible to obtain closed form expressions for its expectation and variance:

$$\mathbb{E}[\log(Y)] = \log(x_m) + \frac{1}{\alpha}$$

$$\mathbb{E}[\log(Y)^2] = (\log x_m)^2 + \frac{2 \log x_m}{\alpha} + \frac{2}{\alpha^2}$$

Using these facts, obtain the element $(2, 2)$ of the Fisher information matrix.

(f) How you would obtain a 95% confidence interval for β_1?