R Individual Assignment – Semester 1, 2022/23

The *Pareto distribution*, named after the Italian economist Vilfredo Pareto, is a power-law probability model that is used in many types of observable phenomena[1]. This was originally used to describe the distribution of wealth in a society, fitting the trend that a "a large portion of wealth is held by a small fraction of the population"[2].

The goal of this assignment is to fit an appropriate Pareto distribution onto a sample data of annual household incomes for a certain country, and use the distribution to estimate the quantiles of income in that country.

The *probability density function (pdf)* of the Pareto distribution is the real-valued function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x \mid \alpha, \beta) = \begin{cases} \frac{\alpha\beta^\alpha}{x^{\alpha+1}} & x \geq \beta \\ 0 & x < \beta \end{cases} \tag{1}$$

There are two *parameters* of the Pareto distribution:

- The *shape* parameter $\alpha > 0$ (also known as the Pareto index); and
- The *location* parameter $\beta > 0$ (also known as the scale parameter).

QUESTION 1

Write a function called `pareto_pdf(x, alpha, beta)` that takes three arguments (the value of x, the shape parameter α, and the location parameter β) and returns the value of $f(x \mid \alpha, \beta)$. This function must work correctly for all values of x, and for *valid* values of α and β (return an error otherwise).

Let $X := \{X_1, \ldots, X_n\}$ be an independent *random sample* from an assumed Pareto distribution with unknown shape and scale parameters α and β. We can ***estimate*** these values by using the *maximum likelihood estimation* technique.

First, we define the *deviance function* $D(\alpha, \beta \mid X)$ mapping the two parameter values to some real value, i.e. $\mathbb{R}^+ \times \mathbb{R}^+ \mapsto \mathbb{R}$, as

$$D(\alpha, \beta \mid X) = -2 \sum_{i=1}^{n} \log f(X_i \mid \alpha, \beta),$$

where f is as previously defined in (1). Note that the deviance function returns the value of the sum of the natural logarithm of the pdf (multiplied by -2) using each of the data values X, evaluated at input parameter values α and β.

QUESTION 2

Write a function `pareto_dev(alpha, beta, x)` that takes three arguments (the shape parameter α, the scale parameter β, and the data vector X) and returns the value of $D(\alpha, \beta \mid X)$. Note that this function should not be vectorised over `alpha` nor `beta`.

The maximum likelihood estimators (MLE) $\hat{\alpha}$ and $\hat{\beta}$ of the parameters satisfy

$$\hat{\alpha} = \arg\min_{\alpha} D(\alpha, \hat{\beta} \mid X) \qquad \text{and} \qquad \hat{\beta} = \min\{X_1, \ldots, X_n\}.$$

[1] `https://en.wikipedia.org/wiki/Pareto_distribution`

[2] The *Pareto principle* or "80-20" rule, stating that 80% of outcomes are due to 20% of causes, was named in honour of Pareto. Empirical evidence suggests that this 80-20 rule fits a wide range of cases in nature as well as in human activities.

In other words, the location parameter β may be estimated using the minimum value of the sample X; while the shape parameter α may be estimated by using the value of α (given $\hat{\beta}$ and X) that <u>minimises</u> the deviance function.

There are several approaches to finding $\hat{\alpha}$:

- Compute (by hand) the partial derivative of D with respect to α, and determine the closed-form value of α where this partial derivative is zero.
- Use of a built-in optimiser such as `optim()` in R. If you use this approach, you should use `method = "L-BFGS-B"` and supply an appropriate `upper` and/or `lower` bound. See `??optim` for more information.

QUESTION 3

Download the data file for this assignment from Canvas. Load this data file in R, and create a vector `X` containing the data points. Code the MLE of $\hat{\alpha}$ and $\hat{\beta}$ based on the data `X`, and make sure there are two objects (in the Global environment) called `alpha_hat` and `beta_hat` corresponding to the MLEs respectively.

Next, we introduce the *cumulative distribution function (cdf)* $F : \mathbb{R} \to [0, 1]$ defined by

$$F(x \mid \alpha, \beta) = \Pr(X \leq x) = \int_{-\infty}^{x} f(u \mid \alpha, \beta)\, \mathrm{d}u.$$

In other words, the cdf returns the area under the pdf f up to the point $x \in \mathbb{R}$.

QUESTION 4

Write a function called `pareto_cdf(x, alpha, beta)` that takes the usual three arguments and returns the value of the cdf $F(x \mid \alpha, \beta)$. You may compute F by hand, or if you prefer, use the `integrate()` function in R.

Specifically, we are interested in five points $(q_1, q_2, q_3, q_4, q_5)$ of the Pareto distribution:

$$F(q_i \mid \alpha, \beta) = \begin{cases} 0.05 & i = 1 \text{ (the 5th percentile)} \\ 0.25 & i = 2 \text{ (the lower quartile)} \\ 0.50 & i = 3 \text{ (the median)} \\ 0.75 & i = 4 \text{ (the upper quartile)} \\ 0.95 & i = 5 \text{ (the 95th percentile)} \end{cases}$$

QUESTION 5

Find the values $(q_1, q_2, q_3, q_4, q_5)$ of the Pareto distribution with parameter values equal to the MLEs that you obtained in Question 3 above, using the method described below.

- Create a vector `x_vals` of length `B` starting from `xmin` and ending at `xmax`.
- Using loops, create another vector `cdf_vals` containing values of F(`x_vals`).
- The value of q_i can be *subsetted* using information from `x_vals` and `cdf_vals` (e.g. the value of q_3 is the `x_vals` value at the index where `cdf_vals` is *closest* to 0.5). Assign a vector named `qvals` containing the q_i values.
- You are free to choose the values `B`, `xmin` and `xmax`; but note that these values will affect the quality of the estimated q_is.