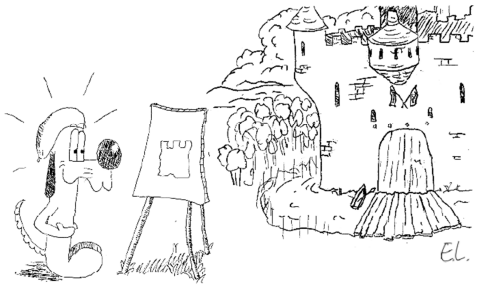
Performance Evaluation Exercises

Model Fitting 2

With Solutions Jean-Yves Le Boudec, Spring 2021



1. We have a data setx_i, $i = 1 : n$ with n large, for which we found that the hazard rate $\lambda(x)$ becomes small when x is large. From this observation, which of the following distributions could be envisioned to model the data ?

   (a) ☐ A normal distribution

   (b) ☒ A Weibull distribution with shape parameter $0 < c < 1$

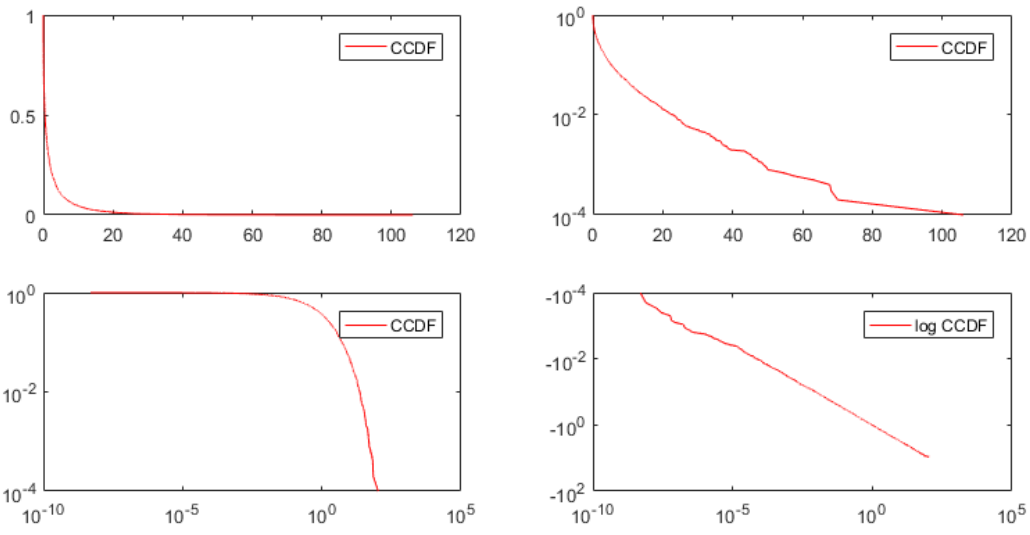
   (c) ☐ An exponential distribution

   (d) ☐ A Weibull distribution with shape parameter $c > 1$

   (e) ☒ A Pareto distribution with index $0 < p < 2$

   (f) ☒ A Pareto distribution with index $p \geq 2$

2. (Continued) We plot the survival function of the data set in the previous question (i.e. the complementary CDF (CCDF)). We obtain the following plots in various scales.



   Which distribution do you propose to model this data set ?

   (a) ☐ A Pareto distribution with $0 < p < 2$

   (b) ☐ A Pareto distribution with $2 \leq p$

   (c) ☒ A Weibull distribution

Solution. The log of the empirical CCDF has a power-law in log-log scale, i.e. there is a relation of the type

$$\log\Big(-\log\hat{F}(x)\Big) \approx a\log x + b$$

where $\hat{F}$ is the complementary CDF (note that $\log\hat{F}$ is negative so the log log plot uses the log of $\Big(-\log\hat{F}(x)\Big)$). This relation should also hold for the theoretical CCDF $\bar{F}$. For Pareto we have

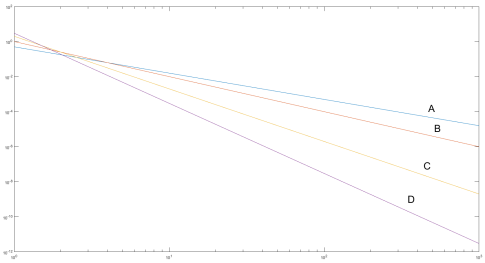
$$\log\big(-\log\bar{F}(x)\big) = \log\left(-\log\left(\frac{1}{x^p}\right)\right) = \log p + \log\log x$$

so this does not correspond. For Weibull we have

$$\log\big(-\log\bar{F}(x)\big) = \log\left(-\log e^{-(x^c)}\right) = \log x^c = c\log x$$

which does correspond.

3. Find the index of each of the standard Pareto PDFs shown in the figure



(a) ☒ $A = 0.5, B = 1, C = 2, D = 3$

(b) ☐ $A = 3, B = 2, C = 1, D = 0.5$

(c) ☐ $A = 0.5, B = 2, C = 1, D = 3$

(d) ☐ $A = 1, B = 2, C = 3, D = 0.5$

Solution. The decay is faster when the exponent is larger. Note that we have a log-log plot, hence the straight lines, indicative of power laws.

4. The complementary CDF of a Pareto distribution follows a power law...

(a) ☒ True

(b) ☐ False

(c) ☐ It depends on the index p

5. A Pareto distribution is heavy tailed (i.e. with infinite variance)...

(a) ☐ True

(b) ☐ False

(c) ☒ It depends on the index p

Solution. It is heavy tailed for $0 < p < 2$ and not heavy tailed (finite variance) for $p \geq 2$.

6. For a Pareto distribution, the hazard rate $\lambda(t)$ is such that $\lim_{t\to\infty}\lambda(t) = 0$.

(a) ☒ True

(b) ☐ False

(c) ☐ It depends on the index p

Solution. The hazard rate is $\lambda(t) = \frac{f(t)}{1-F(t)} = \frac{p}{t}$ and this holds for all p. The hazard rate vanishes for large t, i.e. Pareto(p) is fat-tailed for any value of p (but is heavy tailed only for $p < 2$).

7. The distribution of the sum of n iid random variables with heavy tail and index $p < 2$, for large n, is approximately...

   (a) ☐ Normal
   (b) ☒ Stable with same index p
   (c) ☐ Stable but not necessarily with same index p
   (d) ☐ Poisson
   (e) ☐ It depends on p

8. X is a random variable with distribution standard Pareto with index $p > 0$. The distribution of $\log(X)$ is ...

   (a) ☐ Normal
   (b) ☐ Stable with same index p
   (c) ☐ Stable but not necessarily with same index p
   (d) ☐ Poisson with rate $\lambda = \frac{1}{p}$
   (e) ☐ Poisson with rate $\lambda = p$
   (f) ☒ Exponential with rate $\lambda = p$
   (g) ☐ Exponential with rate $\lambda = \frac{1}{p}$
   (h) ☐ Lognormal

Solution. Method 1. Whenever $Y = \varphi(X)$ and φ is increasing, the change of PDF is given by $f_X(x) = \varphi'(x) f_Y(\varphi(x))$. Apply this to $Y = \log(X)$, i.e. $\varphi = \log$ and obtain

$$f_X(x) = \frac{1}{x} f_Y(\log(x))$$

Let $y = \log(x)$, i.e. $x = e^y$, in the previous formula and obtain

$$f_Y(y) = x f_X(x) = x \frac{p}{x^{p+1}} = \frac{p}{x^p} = \frac{p}{e^{py}} = pe^{-py}$$

which shows that by $Y \sim \text{Exp}(p)$.

Method 2. By CDF inversion, a sample of X is obtained by $X = \frac{1}{U^{\frac{1}{p}}}$ where $U \sim$Unif$(0,1)$. Thus a sample of $\log(X)$ is obtained by $\log(X) = -\frac{\log(U)}{p}$.

Similarly, we know that a sample Y of Exp(λ) is obtained by $Y = -\frac{\log(U)}{\lambda}$. Thus the distribution of $\log(X)$ is Exp(p).

Method 3. Let $Y = \log(X)$. For any test function φ we have

$$\mathbb{E}\left(\varphi(X)\right) = \mathbb{E}\left(\varphi(\log(X))\right) = \int_1^{+\infty} \varphi(\log(x)) \frac{p}{x^{p+1}} dx$$

Do the change of variable $y = \log(x)$ in the integral and obtain $dy = \frac{dx}{x}$ hence

$$\mathbb{E}\left(\varphi(X)\right) = \int_0^{+\infty} \varphi(y) \frac{p}{e^{py}} dy = \int_0^{+\infty} \varphi(y) p e^{-py} dy$$

which shows that the PDF of Y is $f_Y(y) = pe^{-py}$ for $y \geq 0$, i.e. $Y \sim \text{Exp}(p)$.

9. The distribution of the sum of n iid random variables with finite variance, for large n, is approximately...

   (a) ☒ Normal

   (b) ☐ Stable

   (c) ☐ Poisson

   (d) ☐ It depends

   Solution. This is the central limit theorem.

10. X_i is an iid sequence with PDF $f_{X_i}(x) = \frac{2}{\pi(1+x^2)}\mathbf{1}_{\{x\geq 0\}}$ (one-sided Cauchy). When n is large, what can you say about $Y = X_1 + ... + X_n$?

    (a) ☐ It is approximately gaussian

    (b) ☐ It is approximately stable with index $p = 2$

    (c) ☒ It is approximately stable with index $p = 1$

    Solution. The distribution of X_i is heavy-tailed with index $p = 1$. By aggregation, Y is approximately stable with $p = 1$ (and is also heavy tailed with $p = 1$).

11. How do you generate a sample of the standard Weibull distribution with shape parameter c ?

    Solution. By CDF inversion:

    $$1 - F(x) = e^{-(x^c)}$$

    hence we can draw $U \sim \text{Unif}(0, 1)$ and solve for X in

    $$e^{-(X^c)} = 1 - U$$

    which gives

    $$X = (-\log(1 - U))^{\frac{1}{c}}$$

    Observe that U and $1 - U$ have the same distribution so we can also use the following formula:

    $$X = (-\log U)^{\frac{1}{c}}$$

12. What are the models and the null hypothesis of a Jarque-Bera test ? Give a formula to compute the $p-$value when the data is $x_1, ..., x_n$ and n is large.

    Solution. The model is $X_1, ...X_n$ is iid from some distribution. H_0 is : this distribution is N_{μ,σ^2} for some μ and $\sigma > 0$.

    The test statistic $T(x)$ is

    $$\begin{aligned} T(x) &= \frac{n}{6}\left(\hat{\gamma}_1^2 + \frac{1}{4}\hat{\gamma}_2^2\right) \\ \hat{\gamma}_1 &= \frac{m_3}{\hat{\sigma}^3} \\ \hat{\gamma}_2 &= \frac{m_4}{\hat{\sigma}^4} - 3 \end{aligned}$$

where $\hat{\sigma}^2$ is the sample variance, m_3 is the estimate of the third moment and m_4 of the 4th moment:

$$\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 \\
m_3 &= \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^3 \\
m_4 &= \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^4
\end{aligned}$$

Under H_0, T has a distribution χ^2_2. Further, the theoretical values of skewness and kurtosis indices γ_1 and γ_2 are 0, so under H_0, T should be small. The rejection region is of the form $T(x) > c$ for some constant c, therefore the $p-$value is

$$p^* = \mathbb{P}\left(T(X) > T(x)\right) \approx 1 - \chi^2_2(T(x))$$

where $\chi^2_2()$ is the CDF of the χ^2_2 distribution.

13. (a) What is the complementary CDF of a Pareto distribution with index p rescaled by a factor $s > 0$?

    (b) What is the distribution of a censored standard Pareto random variable, more precisely, the conditional distribution of a standard Pareto random variable X given that $X > a$?

    (c) A data set $X_1, ...X_n$ is assumed to be iid from a censored standard Pareto with index p and censoring parameter a. Write the formula for the maximum likelihood estimation of p and a.

Solution.

(a) The re-scaled distribution is that of sX where $X \sim$ Pareto(p). Thus the CCDF is

$$\begin{aligned}
1 - F(x) \overset{\text{def}}{=} \mathbb{P}\left(sX > x\right) &= \mathbb{P}\left(X > \frac{x}{s}\right) \\
&= \left(\frac{s}{x}\right)^p \text{ for } x \geq s \\
&= 1 \text{ for } x \leq s
\end{aligned}$$

(b) The censored distribution has a complementary CDF given by

$$\begin{aligned}
1 - F(x) \overset{\text{def}}{=} \mathbb{P}(X > x | X > a) &= \frac{1 - F(x)}{1 - F(a)} \text{ for } x \geq a \\
&= 1 \text{ for } x \leq a
\end{aligned}$$

When $X \sim$ Pareto(p), this gives

$$1 - F(x) = \left(\frac{a}{x}\right)^p \mathbf{1}_{\{x \geq a\}}$$

Which is the CCDF of the re-scaled Pareto distribution with scale a. For Pareto, re-scaling or censoring are equivalent ! But this is special: for most other distributions, censoring changes the shape !

(c) Since censored Pareto is the same as rescaled Pareto, the model is

$$X_i \sim \frac{1}{a} \text{ iid Pareto}(p)$$

for some $a > 0$ and $p > 0$. The PDF of rescaled Pareto is

$$f(x) = \frac{1}{a} \frac{p}{\left(\frac{x}{a}\right)^{p+1}} \mathbf{1}_{\{\frac{x}{a} \geq 1\}} = \frac{pa^p}{x^{p+1}} \mathbf{1}_{\{x \geq a\}}$$

The log-likelihood of an observation $X_1 ... X_n$ is thus

$$\begin{aligned} \ell(a,p) &= n \log p + np \log a - (p+1) \sum_{i=1} \log x_i \text{ if } a \leq x_i \text{ for all } i \\ &= -\infty \text{ else} \end{aligned}$$

ℓ is increasing with a and is thus maximized when $a = \hat{a} = \min(x_i)$. To obtain $\hat{p}$ we compute the derivative

$$\frac{\partial \ell}{\partial p} = \frac{n}{p} + n \log a - \sum_{i=1}^{n} \log x_i$$

we see that there is a maximum when $\frac{\partial \ell}{\partial p} = 0$ which occurs when

$$\frac{1}{p} = \frac{1}{n} \sum_{i=1}^{n} \log x_i - \log a$$

In summary, the MLE is

$$\begin{aligned} \hat{a} &= \min_{i=1:n} x_i \\ \hat{p} &= \left(\frac{1}{n} \sum_{i=1}^{n} \log x_i - \log \left(\min_{i=1:n} x_i \right) \right)^{-1} \end{aligned}$$