Input Modeling II- Fit Distribution Via R

W. M. Song 桑慧敏
Tsing Hua Univ. 清華大學

2015.11.11

Install R, RStudio, and JDK

Install R, RStudio, and JDK

- R-project
  URL: http://www.r-project.org/
- RStudio
  URL: http://www.rstudio.com/
- JDK8:
  To produce an open-source reference implementation of the Java SE 8 Platform Specification defined by JSR 337 in the Java Community Process.
  URL: http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html

- Note: URL (Uniform Resource Locator), 俗稱網頁地址 (網址)

Interface of RStudio: 4 areas

Draft

You can type all drafts here

Global Environment

we can see data and value types here

Console (控制台)

R commands are implemented here

Plots, Packages, ...

Files, Plots, Packages, Help, and Viewers

Install Packages

Install Packages "xlsx"

- Goal: we can read file.xlsx from MSExcel
- Install package (Open Rstudio, Click Packages, Click Install)
- Install xlsx package (Type "xlsx" in text box and click install)
- Call Library in RStudio (Type library ("xlsx") in Console)

Install Packages "fitdistrplus"

- Goal: we can use "fitdist" and "ks.test" in R
- Install package (Open Rstudio, Click Packages, Click Install)
- Install fitdistrplus package (Type "fitdistrplus" in text box and click install)
- Call Library in RStudio (Type library ("fitdistrplus") in Console)

Keywords about "Fit Distributions" in R

- Estimate the parameters:fitdist
- Goodness of fit test:ks.test
- Fit distribution package in R: Package 'fitdistrplus'
- How to read xlsx file to R:

  DataE=read.xlsx("D:/Path1/Path2/Arrival_Data.xlsx",1)
  DataE=read.xlsx("D:\\ Path1\\Path2\\Arrival_Data.xlsx",1)
  data=DataE[, 1] (Notice. "[ ]", instead of "( )")
  (Note: 1 denotes for Excel sheet 1)

- "data" is the file name in R containing the same data in DataE

Estimate the parameters

fitdist(data, "beta", "mle")	beta(shape1, shape2)
fitdist(data, "gamma", "mle")	gamma(shape, scale)
fitdist(data, "weibull", "mle")	Weibull(shape, scale)
fitdist(data, "exp", "mle")	exp(beta)
fitdist(data, "norm", "mle")	norm(mean, sd)
fitdist(data, "lognorm", "mle")	lognorm(mean, sd)

- Q: To fit "beta", we need to do some data transformation. See next page.

Fit "Beta (0,1)" Distribution

How to fit "Beta (0,1)" to data X

- Q: What if some values of X do not in (0, 1), range of Beta Dist.?
- A: Via transformation, we have $X = a + bY$, where
  Y follows Beta distribution,
  $a = \min(X) - 10^{-6}$, and
  $b = (\max(X) - \min(X) + (10^{-6}))/(1 - (10^{-6}))$.

Explanation

- Since we can not set min of Y to be 0, and the max of Y to be 1, we set min(Y)$= 10^{-6}$, and max(Y)$= 1 - 10^{-6}$.
- We can translate X_i to Y_i, $i = 1, 2, ..., n$ such that all $Y_i \in (0, 1)$.
- $Y = [X - \min(X) + 10^{-6}]/b$, where
  $b = [\max(X) - \min(X) + (10^{-6})]/[1 - (10^{-6})]$.

Goodness of fit test

```
ks.test(data, "pbeta", shape1, shape2)
ks.test(data, "pgamma", shape, scale)
ks.test(data, "pweibull", shape, scale)
ks.test(data, "pexp", scale)
ks.test(data, "pnorm", mean, sd)
ks.test(data, "plognorm", logmean, logsd)
```

Results From Arrival_Data

Name of fitted Dist	Parameter 1	Parameter 2	p-value
Beta(shape1,shape2)	0.722271	3.061547	0.5286
Gamma(shape,scale)	0.8857522	0.174542	8.69E-10
Weibull(shape,scale)	0.9487846	4.9641585	7.302E-10
Exp(beta)	0.1970715		2.67E-14
Norm(mean,sd)	5.0743	4.776929	2.20E-16
LogNorm(mean,sd)	0.9628464	1.4446337	2.20E-16

- Q: Which distribution fit the best?
- A: The larger the p value, the better the fit. (See Testing Hypothesis, Output Analysis II)

Homework:Regis_Data

- Fill the cell in the following table
- Choose the best fitted distribution

Name of fitted Dist	Parameter 1	Parameter 2	p-value
Beta(shape1,shape2)			
Gamma(shape, scale)			
Weibull(shape,scale)			
Exp(beta)			
Norm(mean,sd)			
LogNorm(mean,sd)			