International Journal of Mathematical Archive-5(1), 2014, 118-124
Available online through www.ijma.info ISSN 2229 – 5046

USING EASY FIT SOFTWARE FOR GOODNESS-OF-FIT TEST AND DATA GENERATION

Hossein Mehrannia*[1] & Alireza Pakgohar[2]

[1]Department of Mathematics, Behbahan Branch, Islamic Azad University, Behbahan, Iran, E-mail: hmerannia210@gmail.com

[2]Science assistance, Statistics department, Payam-e Nour University, P.O. Box 19395-3697 Tehran, Iran, E-mail: a_pakgohar@pnu.ac.ir

(Received on: 25-12-13; Revised & Accepted on: 16-01-14)

ABSTRACT

Data Fitting process involves using certain statistical techniques which allow us to estimate fitness parameters in accordance to data sample. One advantage of using software to fit the data and interpreting probability data is that they are able to automatically fit data with a variety of known distribution patterns simultaneously. These methods are preferred especially in cases which you have very little or no information about the base distribution pattern in data, and desire to find the best distribution type. EasyFit is a data analyzer and simulation software which allows us to fit probabilistic distributions to given data samples, simulate them, choose the best fitting sample, and implement the results of analysis to take better decisions. This software can be used as a Windows compatible program, and also as an add-on to Excel spread sheets. This study reviews performance, the abilities, and imbedded statistical tools of this software.

Keywords: *Easy Fit software, Goodness-of-Fit test, Kolmogorov-Smirinov, Chi squared, Anderson-Darling.*

1. INTRODUCTION

When a certain distribution is chosen as data distribution, it is expected to fit suitably with data, so we are ready to practically fit those data with the distributions.

Data fitting process includes using certain statistical techniques which allow us estimate fitness parameters based on sample data. The distribution fitting software can be very useful in this sense. Clearly this program will use the methods of parameter estimation on the best known distributions, though saving your time and letting you to concentrate on data analysis task. Since researchers sometimes want to fit several different distributions to the data simultaneously, they need to estimate parameters of each distribution separately. Input to the data fitting software usually includes:

- Your data in any accepted format
- Distributions which you want to fit with them

Fitting choices

Output results of the fit would include:

- Graphs related to your raw input data
- Distribution or improved fitting parameters
- Graphs of fitted distribution
- Additional graphs and tables which help you choosing best fit for your data.

1.1 Automatic distribution fitting

One of the advantages of using computer software to fit the data in probabilistic data analysis is that they have the ability to automatically fit data to a large number of distributions simultaneously. This working method is preferred in cases which you have little or no information about base distributions existing in data, and want to find the general type of distribution.

Corresponding author: Hossein Mehrannia[1]*
[1]Department of Mathematics, Behbahan Branch, Islamic Azad University, Behbahan, Iran, E-mail: hmerannia210@gmail.com

2.1 choosing the best fit for distribution

As the distributions are fitted, you can compare them to each other and choose the best given fit model. There are several statistical techniques and tools which can help you doing this task. These tools are usually integrated into the distribution-fitting software; and are used as various types of graphs and tables which show the distribution along with estimated parameters.

Distribution Graphs

The distribution graphs enable you to:

- Visually evaluate the extent of Goodness-of-Fitting for a certain distribution.
- Compare various fitted models together.

Some of graphs (such as histogram), however, show your very own data along with fitted distribution simultaneously:

- Probability Density Function (PDF) graph
- Cumulated Distribution Function (CDF) graph

Following graphs only show the fitted distribution:

- p-p Graph
- Q_Q Graph
- Probability Difference (PD) graph

Each type of graph has its particular meaning and interpretation. The data fitting program will show you one or more graph/s of fitted distribution according to the choices you made. In manual fitting mode, the graphs would be automatically updated if you modify or change the distribution parameters; this gives a more interactive nature to the fitting process.

2. STATISTICAL THEORIES

In this section we try to present a general review of statistical theories and techniques in data fitting and Goodness-of-Fit test field. We would have an overview on statistics and related graphs as well.

1.2 Goodness-of-Fit tests

Goodness-of-Fit tests, as suggested by their very name, can be used to determine whether a certain distribution is fitted properly to the data or not. Calculating statistics of Goodness-of-Fit also helps to rank the fitted distributions according to quality of fit over the raw data. This particular characteristic feature of the software is very useful for comparing fitted models to each other.

Most used Goodness-of-Fit tests include Kolmogorov-Smirinov, Anderson-Darling, and Chi squared tests. Usage logic is all similar for these tests. However, they differ in practical method (and type of usage). Kolmogorov-Smirinov test can be named as the most used Goodness-of-Fit test.

1.3 Implementing the chosen distribution

The higher end for doing analysis is that you obtain information which help you take best informed decisions under uncertainty conditions. The required information can be obtained using best fitted distribution. This distribution is a model of stochastic process which we face in real world.

1.4 General application examples

Some applications of statistical distributions include:

Event probability calculation

- Estimations
- Statistical indices or statistics calculation

These calculations can be done using relevant functions for desired distributions; including Cumulative Distribution Function (CDF) and inverse CDF, and Hazard Function (HF).

One of the most implications of these functions includes calculating likelihoods: You specify a good (desired) result in a usual data analysis; and calculate the likelihood of that result to take place. It is worthwhile to take the decision leading to that result only if the level of likelihood is high enough. On the other hand, you should consider the opposite-side result if the likelihood is low.

For example, in analyzing normal distribution of service time to clients, the desired result maybe as the following:

- Good result: Being able to serving each client in less than 5 minutes.
- Bad result: Serving each client takes more than 5 minutes of time.

Relevant decisions in this case are:

- Decision (a): Hire no more personnel
- Decision (b): Hire more personnel

The likelihoods can be very simply calculated using cumulated distribution function (CDF) fir chosen distribution. For example CDF(5) indicates the likelihood of obtaining good result from the process. In case the value returned by this function is less than a certain threshold (e.g. <95%), decision of hiring more personnel should be considered as a solution which reduces client serving time and improves customer satisfaction. 90% likelihood would mean that 10% of your customers have to wait more than 5 minutes and may get unsatisfied with services provided by your firm to clients.

In some cases you want to define more than one result:

- Result (a): we are able to serve each client in less than 5 minutes.
- Result (b): The service time is between 5 and 6 minutes for each client.
- Result (c): Serving each client takes more than 6 minutes of time.

These likelihoods can be found in a similar way. The outcome maybe as following:

- Likelihood of result (a) = 90%
- Likelihood of result (b) = 7%
- Likelihood of result (c) = 3%

You may decide not to hire more personnel in this case, since only 3% more clients are receiving services in more than 6 minutes

.

1.5 Estimation

Estimation is a reverse problem for which you need to consider a fixed value as likelihood. For example, suppose you want to estimate the service time for 95% of your clients. The Inverse Cumulated Distribution Function or ICDF can be used on your chosen distribution in this task: ICDF (.95) = 5.5. It is interpreted as: Although 90% of clients are receiving their desired service in less than 5 min. (example above), but 5% of clients will wait 5:30 minutes (30 seconds more) for being served; which is totally acceptable.

1.6 calculating statistics

Calculating statistics can be useful in obtaining a general perspective of the data (note that it is not wise to make decision on sheer bases of statistical figures or indices.) Most useful statistics are:

- Mean (average value)
- Mode (most probable value)

For example, you may find that possibly one client will receive the desired service in about 2 minutes, but there are many others who need more time to receive the service, so the mean service time is about 3 minutes.

1.7 Especial implications

Although probabilistic distributions can be used in fields which work with stochastic data, but other applications exist in certain industries (including statistical science, financial studies, safety engineering, water science, etc…) which help business managers, engineers and scientists to make informed decisions under uncertainty conditions.

3. EASYFIT SOFTWARE

EasyFit © is a data analysis and simulation software which enables us to fit and simulate statistical distributions with sample data, choose the best model, and use the obtained result of analysis to take better decisions. This software can function as a stand-alone windows application or as an add-on for Excel spread sheet.

Prominent features of this program are:

- Supports more than 50 discrete and continuous distributions
- Automatic and manual settings
- Ability to test performed operations
- Bi-modal graphs
- Random number generation
- Integrated help system

1.8 Excel integration

EasyFit program is easily integrated in main menu of Excel and allows you to implement your analysis and simulation in Excel environment.

This software benefits of more than 650 spread sheets in Excel which can facilitate the calculation tasks.

With EasyFit program you can:

- Analyze large datasets (up to 250,000 data points)
- Calculate the graphical statistics
- Organize your data and enter obtained results in your project file.

There are two methods for data fitting: Automatic and Manual.

1.9 Automatic distribution fitting method

Most applied method for fitting data with probabilistic distributions is to use automatic fitting feature of the EasyFit. This feature allows you to fit a large number of statistical distributions to the data in a batch to batch process. This technique is most useful when there is very little or no information available regarding possible statistical distribution of the data.

Data input

You need to import your data to EasyFit in order to analyze them. This can be done through a file, or through clipboard using Copy/Paste commands. Or data could be entered manually.

Choosing particular options of distribution fitting

Before EasyFit automatically fits the distributions to your data, you can determine exactly which distributions you want to be fitted. There are a large number of probabilistic distributions available in EasyFit. You may want to use all or only some of them depending on whether you need additional information about you data or not. You can use the *"Distribution fitting options"* (from *"Options -> fitting"* menu) to specify distributions which you want to fit. There are a few more choices for fitting which can be used, but we address only the basic options in this review:

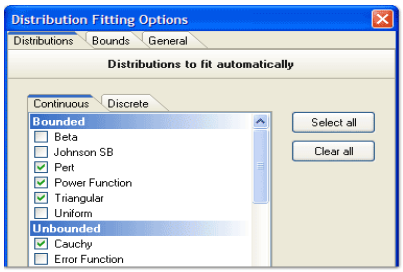


Fig (1): Statistical distribution tab in automatic method

Click on "OK" button to close the dialogue box and choose "*Analyze -> Fit Distribution"* from main menu. EasyFit will ask you to specify your input data, then starts the process of fitting the distributions.

Done the fitting of distributions, EasyFit will rank according to quality level of fitting with data based on Goodness-of-Fit statistics and sorts them out:

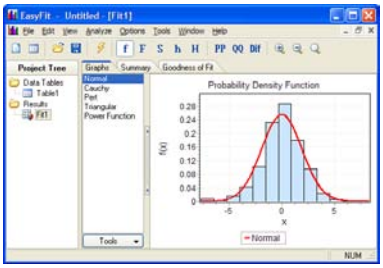


Fig (2): Dialogue box showing fitted distributions graph in automatic mode

Now you can use the distribution graphs and Goodness-of-Fit test results to compare fitted distributions and choose the best model.

1.10 Manual fitting of data

The capability of EasyFit software for manually fitting the data allows you to modify fitted distributions in various forms – you can change distribution parameters, fit more distributions, or delete them from the list of fitted models. To fit a new distribution to your data, right click on distributions list and choose "*Add Distribution*".

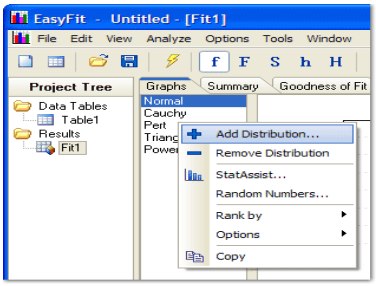


Fig (3): Statistical distribution list in manual mode

EasyFit will show a list of available distributions which helps you easily find and specify your desired model:

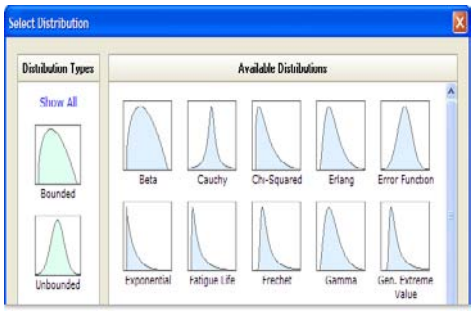


Fig (4): Distribution graphs dialogue box in manual mode

As a distribution is specified, EasyFit will estimate its parameters and adds it to the list of fitted models. Estimated parameters are shown in a separate box which can be used for modifying parameter values.

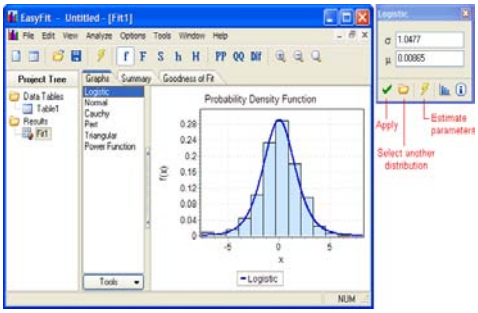


Fig (5): Fitted distributions graphic dialogue box in manual mode

EasyFit software will plot the graphs connected to related distribution if you click on *"Apply"* button and modifies the Goodness-of-Fit reports related to parameters specified by you.

To delete one or more distributions from this list, select them using your mouse and right-click on them, then choose "remove distribution" option.

4. RUNNING SOFTWARE

New we are almost ready to run the program. This section answers two essential questions which could be included in list of important questions among users.

4.1 Choosing the best fitted model using Goodness-of-Fit tess

Suppose you have fitted a variety of distributions with data, and now you need to determine the most credible model. How you compare the fitted distributions? How you can know whether a distribution has been fitted suitably with data? Goodness-of-Fit (GOF) tests can help you answer these and several other questions.

4.2 How do goodness-of-Fit tests work?

Basic idea behind goodness of fit tests is that the "*distance*" between data and the distribution under test is measured; and is compared to a certain threshold value. If this distance (called "*test statistics*") is lower than threshold (critical value), distribution is considered as "good".

The logic of performing more than one type of test is similar. These tests however, are different in test statistics measurement method and critical value calculations. The test statistics are usually defined in form of a function of sample data and cumulated (fitted) distribution function.

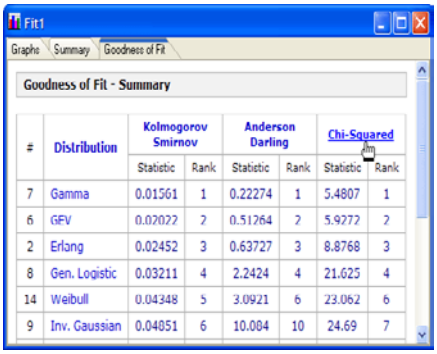
The critical values depend on sample size and significance level which is chosen. The significant level is the probability of rejecting a fitted distribution (as a bad distribution) while it is in fact a good one. Significance level is usually shown by α (alpha); and most used values for it are 0.05 and 0.01. For example, the probability of (wrongly) rejection for a good fit is 5% if the value of 0.05 is specified as significance level of fitting tests.

In EasyFit, you can use almost all the Goodness-Of-Fit tests including Kolmogorov-Smirinov, Anderson-Darling, and Chi square tests. When the distributions are fitted, EasyFit will generate a report of goodness-of-fit values which includes calculated test statistics and critical values for various significance levels:

Fig (5): Output window – Goodness-Of-Fit statistics

Goodness-Of-Fit reports can be used to determine whether a certain probabilistic distribution is fit properly to data or not.

We compare the process of fitting for several kinds of distribution. Since Goodness-Of-Fit statistics are in form of distance between data and fitted distributions, clearly the distribution with minimum statistics value has been best fitted with data. Based on this fact, EasyFit will attribute a ranking number to each distribution (1-the best model, 2-best model after the first one … etc). This allows you to select the most reliable model easily.



#	Distribution	Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
		Statistic	Rank	Statistic	Rank	Statistic	Rank
7	Gamma	0.01561	1	0.22274	1	5.4807	1
6	GEV	0.02022	2	0.51764	2	5.9272	2
2	Erlang	0.02452	3	0.63727	3	8.8768	3
8	Gen. Logistic	0.03211	4	2.2424	4	21.625	4
14	Weibull	0.04348	5	3.0921	6	23.062	6
9	Inv. Gaussian	0.04851	6	10.084	10	24.69	7

Fig (6): Output window – GoF statistics comparison

To sort out the fitted distributions based on a GoF test, just click on the name of that test. There is a similar option on graphs page as well; you can access that option by right clicking on the list of distributions:

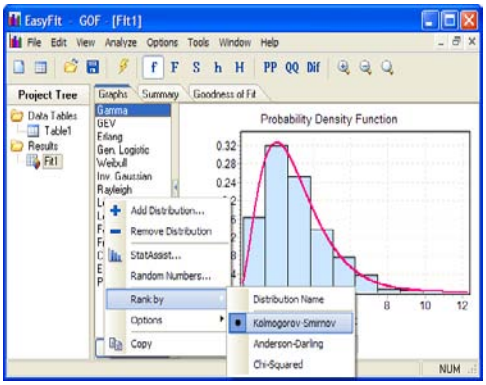


Fig (7): fit graph output window for GoF test

In above example, distributions are sorted based on results from Kolmogorov-Smirnov test statistics, and the best fitted distribution (Gamma) is shown at top of the list.

5. CONCLUSION

Goodness-of-Fit tests can be used to compare fitted distributions, select a model, and determine how good the distribution is fitted to the data. EasyFit software generates reciprocal reports which facilitate achieving a general perspective over fitted distributions as well as evaluating the level of fit goodness for certain models at various significance levels.

BIBLIOGRAPHY

1. Kolmogorov-Smirnov test ,http://mirzadeh.blogfa.com
2. Non-parametric tests, http://ehsanamar.persiangig.com/document/nemoo%20soalne/mb3.pdf
3. Goodness-of-Fit statistics; Alireza Pakgohar,http://pakgohar.blogfa.com/post-513.aspx
4. Anderson-Darling statistics, ne.aidepikiw.gro/ikiw/nosrednA-gnilraD
5. John Netter, Wasserman, Vitmore; “Applied Statistics” vol. 2, Translated by Ali Amidi, Nashr Daneshgahi Center, 2005.
6. EasyFit software manual, ver. 5.4,www.mathwave.com/downloads
7. Rahmani, Tayebe; Pakgohar, Alireza; Modeling and Optimization: M.S. Statistics project, Payam-e Nour University, 2009.

Source of support: Nil, Conflict of interest: None Declared

