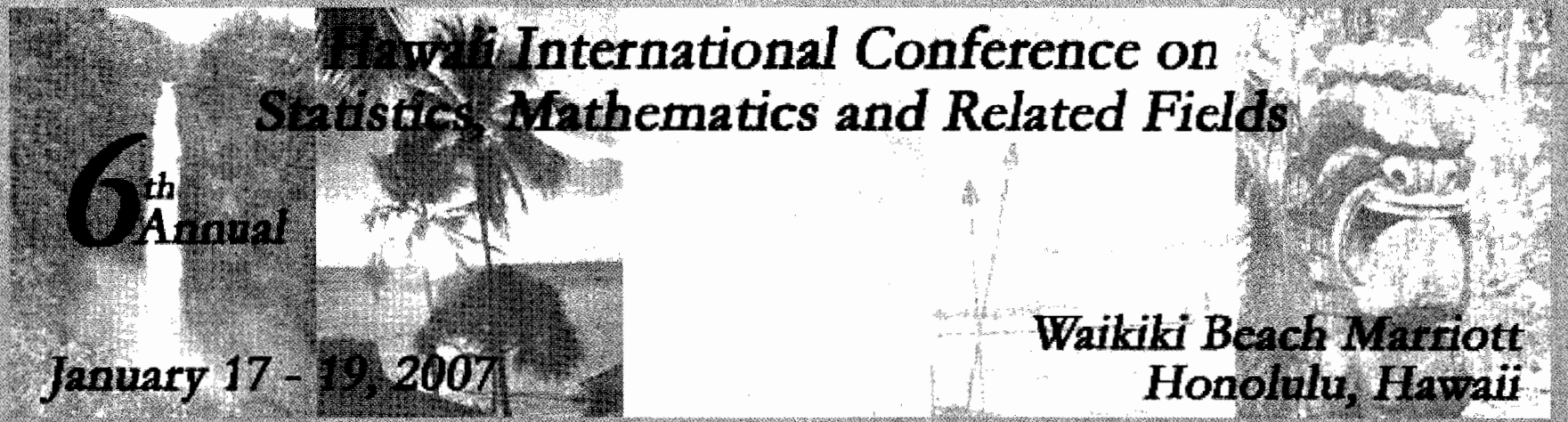


Hawaii International Conference on Statistics, Mathematics and Related Fields

6th Annual

January 17 - 19, 2007

Waikiki Beach Marriott
Honolulu, Hawaii

2007 Conference Proceedings

We would like to thank all those who attended the 2007 Hawaii International Conference on Statistics, Mathematics and Related Fields. We look forward to seeing you at the 7th Annual Conference to be held in 2008. Please check the website this spring for dates and further details.

To search for a specific paper presented, or to browse all of the proceedings, please click the appropriate button on the right.

Papers by Author Name

Papers by Topic Area

Browse Proceedings

Hawaii International Conference on Statistics, Mathematics and Related Fields
PO Box 75023, Honolulu, HI 96836
808-946-9927 (phone) / 808-947-2420 (fax)
statistics@hicstatistics.org -- http://www.hicstatistics.org

ISSN #: 1550-3747

To view a specific paper, click on page number in the last column				
LastName	**FirstName**	**Topic**	**Title**	**Page**
Hagiwara	Katsuyuki	Statistical Modeling	Estimation of the Expected Prediction Error of Orthogonal Regression with Variable Components	426
Hammarstedt	Mats	Cross-Disciplinary Areas of Mathematics	Immigrants' Relative Earnings in Sweden – a Quantile Regression Approach	432
Haque	Shirin	Fractals	A Mathematical Modeling of the Relationship of the Subjective Experience of Past, Present and Future Psychological Events	1319
Hein	Derek W.	Discrete Mathematics	The ?–Design Conjecture: A Survey	462
Hepler	Z.	Cross-Disciplinary Areas of Statistics/ Environmental Health Statistics	PM(2.5) Concentrations from a Maryland Barrier Island Resort Community, 2004 - 2005	1194
Higashi	Shohei	Management Science	Suggestion that the Strategy To Promote the Purchasingat the Same Time in a CD Store	468
Hirose	Hideo	Applied Statistics, Medical Statistics, Epidemiology, Cross-disciplinary areas of Statistics	Estimation for the Number of Fragile Samples in the Trunsored and Truncated Models with Application to the Case Fatality Ratio for the Infectious Diseases	478
Hirose	Hideo	Operations Research	Assessment of the Prediction Accuracy in the Bump Hunting Procedure	491
Hirsch	Christian	Mathematics Education	CPMP-Tools: Emerging Java-based Mathematics Software to Support Student Investigation and Problem Solving	514
Ho	Lun-Pin	Biostatistics	A Decision Analysis Tool – Decision Tree Builder	214
Ho	Lun-Pin	Computational Mathematics	The Implementation and Testing of Gay-Schnabel's Proposal on Systems of Nonlinear Equations	1196
Hrastinski	Stefan	Educational Statistics	E-learning use in Higher Education: Driving Factors, Barriers and Strategies	750
Hromadka, Ph.D.,Ph.D.,Ph. D.	Theodore V.	Statistical Modeling	An Expandable Mathematical Program for Fitting Families of Basis Functions to Data	546
Huang	Ying Sue	Applied Mathematics	Mathematical Models of Delayed Cellular Neutral Networks	522
Ibrahim	N. A.	Computational Statistics	An EM Algorithm for Competing Risks Model	325
Inniss	Tasha R.	Operations Research	Modeling Emergency Evacuation Strategies using Network Models	1223
Inoue	Hiroshi	Financial Mathematics	Option Pricing for which Payoff Depends on Weighted Sums of Prices	528
Inoue	Hiroshi	Financial Mathematics	Multi-period Portfolio Selection Problem with Maximum Absolute Deviation Model	1276
Ishitani	Hiroshi	Statistical Modeling	Estimation of the Expected Prediction Error of Orthogonal Regression with Variable Components	426
Iwabuchi	Takaaki	Business Statistics	A Model of Web Site Browsing Behavior Estimated on Clickstream Data	543
Iyer	Hariharan	Applied Statistics	Impacts of Sources of Variation on Experiments in Random and Mixed Effect Models	258
Jaki	Thomas	Computer Simulation, Mathematical Statistics - Panel	Maximum Kernel Likelihood Estimation	546
Jeong	Hyeok-Je	Number Theory	On the Density of Prime Numbers	547

Title:

AN EXPANDABLE MATHEMATICA PROGRAM FOR FITTING FAMILIES OF BASIS FUNCTIONS TO DATA

Authors:

Theodore V. Hromadka, Ph.D., Ph.D., Ph.D., COL Andrew G. Glen, Ph.D., MAJ Fernando Miguel

Affiliation:

Department of Mathematical Sciences, United States Military Academy

West Point, NY 10996

Emails:

Theodore.Hromadka@usma.edu; Andrew.Glen@usma.edu;

Fernando.Miguel@usma.edu

AN EXPANDABLE MATHEMATICA PROGRAM FOR FITTING FAMILIES OF BASIS FUNCTIONS TO DATA

Miguel,F.D.(1); Hromadka II, T.V.(2); Glen,A.G.(3)

Abstract:

A common problem in statistical analysis of data is to consider fitting functions that minimize some pre-defined measure of goodness-of-fit, such as the well-known least squares residual minimization technique, or a weighted residual minimization technique, among others. Generally, a selection of basis functions are chosen, such as polynomials, or other basis functions, and then all of the basis functions are minimized together as a group according to the selected measure. In this paper, we present a Mathematica program that accomplishes the task of fitting prescribed functions to data, but we build the program to investigate combinations of the basis functions within the family of basis functions as an alternative to fitting the entire family of basis functions simultaneously. Because the procedure is automated, we have the advantage of doing such an effort without a significant investment of time or effort. The Mathematica program is easily expandable to include other families of basis functions or other measures of fit. The program measures the goodness of each set of basis functions used and then lists the measures achieved for each attempt. The program is expandable by downloading the program and doing the relevant programming. It is hoped that this program will be extended in the community to include numerous families of fit functions.

INTRODUCTION

The use of a set of basis functions to form a linear combination with coefficients to be determined by minimizing some best fit measure function to data according to some measure of fit is a process that is well-known and so will not be repeated here (Gallant 1987, among many others.) Generally, a set of basis functions are selected, such as a set of monomials, and then this entire set of basis functions is fitted to the target set of data by a technique such as a weighted Gramm-Schmidt approach or other approach.

In this paper, the various combination of basis functions within a prescribed set of basis functions are used, one at a time, in fitting to the data set. For example, if the set of basis functions is simply {1, x}, then possible trial functions are: a, bx, a+bx, where "a" and "b" are constants to be determined in the data fitting process.

Furthermore, in this paper we present the foundations of a Mathematica program to deliver this approach to data fitting, where the vision of this program is for others to extend the software to include other families of basis functions and to continue examining all possible combinations of basis functions from these families for use as trial functions to be fitted to the data set under study.

THE MATHEMATICA PROGRAM

The Mathematica program is currently based upon use of the Mathematica `FindFit` operation which is internal to the Mathematica library of functions available to the user. We include in the program the usual standard families of basis functions such as polynomials. The various combinations of the basis functions are built from the polynomial family up to the dimension programmed into the notebook. The dimension of this family, for example, can be easily extended. For each combination of these monomials, a new trial function is defined, and the data set is fitted using the Mathematica `FindFit` operation and the resulting measure of fit computed. Once all the combinations of monomials are examined, they are ranked according to goodness of fit by sorting the resulting error residual measures.

Other families are operable in this program. Families of typical interest include trigonometric basis functions including extensions such as: $a\ \mathrm{Sin}(b\ x + c) + d$ which results in several variations just with these four parameters. For example, $a\ \mathrm{Sin}(x)$, $\mathrm{Sin}(b\ x)$, $\mathrm{Sin}(x + c)$, $\mathrm{Sin}(x) + d$, $a\ \mathrm{Sin}(b\ x)$ and so forth. This way, all the diversity within a fitting family is examined. With a competing goal in fitting functions to data being the possibility that simpler fitting functions may be better models of the underpinnings of these data, it is instructive to examine all combinations of basis functions from a prescribed family.

The Mathematica code is available to the reader by contacting the first author of this paper. The Mathematica notebook is easily expandable and it is hoped that others will do so and then share with the author. The proposed extended program software for loading into a web-site is currently under construction for publication on the web.

MATHEMATICA CODE

A portion of the Mathematica code is presented below. The structure of this portion of the code demonstrates the strategy we are currently using. We anticipate changes in almost every aspect of this "shareware" type of program, which is another reason why we chose Mathematica as the base code, not to mention the accuracy levels achieved by use of Mathematica.

User must start with an Excel file that is comma delimited. This program will read in whatever *.csv file you name in the first line from the default folder.
User must enter the first block for the best fitting horizontal line "a". The second block is "bx" and finally the third block is "a + bx". We will continue to add more complex models to this process and display the corresponding SSE of the best fit model. It is much easier to evaluate the entire notebook.

```
data2fit = Import["nan1.csv", "CSV"]
n = First[Dimensions[data2fit]];
maxx = First[data2fit[[n]]];
minx = First[data2fit[[1]]];
plotpoints := ListPlot[data2fit, PlotStyle → PointSize[0.02]];
```

```
{{1, 1.5}, {1, 1.9}, {2, 1.8}, {2, 2.1}, {3, 3},
 {4, 5}, {5, 6.8}, {6, 7}, {7, 9}, {7, 8.9}, {8, 11}, {8, 14}, {9, 17}}
```

1. Model is "a"

```
FitModel[1] = "a";
estreglin1 = FindFit[data2fit, a, {a}, x];
a1 = estreglin1[[1]]; a[1] = a1[[2]];
SSError[1] = Sum[(data2fit[[i, 2]] - a[1])^2, {i, 1, n}]; m[x_, 1] = a[1];
plotmodel[1] := Plot[m[x, 1], {x, minx, maxx}];
```

2. Model is "bx"

```
FitModel[2] = "bx";
estreglin2 = FindFit[data2fit, b x, {b}, x];
b2 = estreglin2[[1]]; b[2] = b2[[2]];
SSError[2] = Sum[(data2fit[[i, 2]] - b[2] data2fit[[i, 1]])^2, {i, 1, n}]; m[x_, 2] = b[2] x;
plotmodel[2] := Plot[m[x, 2], {x, minx, maxx}];
```

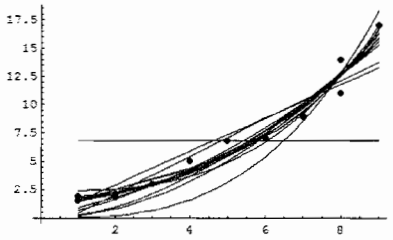
15. Model is $a + bx + cx^2 + dx^3$

```
FitModel[15] = "a +bx + cx^2+ dx^3";
estreglin15 = FindFit[data2fit, a + b x + c x^2 + d x^3, {a, b, c, d}, x];
a15 = estreglin15[[1]]; a[15] = a15[[2]];
b15 = estreglin15[[2]]; b[15] = b15[[2]];
c15 = estreglin15[[3]]; c[15] = c15[[2]];
d15 = estreglin15[[4]]; d[15] = d15[[2]];
SSError[15] =
  Sum[(data2fit[[i, 2]] - (a[15] + b[15] data2fit[[i, 1]] + c[15] (data2fit[[i, 1]])^2 +
        d[15] (data2fit[[i, 1]])^3))^2, {i, 1, n}];
m[x_, 15] = a[15] + b[15] x + c[15] x^2 + d[15] x^3;
plotmodel[15] := Plot[m[x, 15], {x, minx, maxx}];
```

For the polynomial family of distributions, we can break down the different models by order. This code is an exhaustive listing up to the 3rd order. We would expect that when we reach the n^{th} order, we will have brought the SSE down to 0. The purpose of this exhaustive effort, is to show that it may not be necessary to increase the polynomial to the n^{th} order in order to find an appropriate model to fit the data.

```
Show[{plotpoints, plotmodel[1], plotmodel[2], plotmodel[3], plotmodel[4], plotmodel[5],
  plotmodel[6], plotmodel[7], plotmodel[8], plotmodel[9], plotmodel[10], plotmodel[11],
  plotmodel[12], plotmodel[13], plotmodel[14], plotmodel[15]}]
Print["Data to be modeled:"]
Print["x      ", "y"]
data2fit // TableForm
Print["Model                ", "SSE", "Model"]
TableForm[Table[{FitModel[i], SSError[i], m[x, i]}, {i, 1, 15}]]
```

The following output is a combination of 15 models from the polynomial family up to the 3rd order. From this graph, you can conclude that there are a few functions that fit the data in a similar manner. In the Sum Squared Error (SSE) table, you can see that these functions share a similar SSE. We can guess that a further exploration of the 4th order polynomials will yield similar results. Once we have reached the nth polynomial, the SSE will be reduced to zero. Although this is a desired effect, the user can determine the most appropriate model from this family by the frequency at which a small range of SSE is reached. Because our model can extend well below and above the domain of the data, we can graph our model to illustrate extrapolated values. It should become clear that the nth order polynomial with a SSE of zero is likely inappropriate because of the behavior of the predicted values when the independent variable is outside the domain of the original data.



Out[286]//TableForm=

Model	SSE	Model with Parameters
a	299.652	6.84615
bx	31.9619	1.47519 x
a + bx	27.0441	-1.24921 + 1.67047 x
cx^2	21.9511	0.201497 x^2
bx + cx^2	14.6455	0.590312 x + 0.123435 x^2
a + cx^2	11.0295	1.40077 + 0.175657 x^2
a + bx + cx^2	10.6585	1.92982 - 0.303325 x + 0.206009 x^2
dx^3	51.3372	0.0251279 x^3
a + dx^3	12.0762	2.39316 + 0.0200377 x^3
bx + dx^3	11.1631	0.88235 x + 0.0109357 x^3
a + bx + dx^3	9.04134	1.1667 + 0.508571 x + 0.0144662 x^3
cx^2 + dx^3	20.3928	0.259 x^2 - 0.00737106 x^3
bx +cx^2+ dx^3	8.14004	1.68199 x - 0.279428 x^2 + 0.0331362 x^3
a +cx^2+ dx^3	9.78238	1.79291 + 0.101621 x^2 + 0.00856317 x^3
a +bx + cx^2+ dx^3	8.11931	-0.240448 + 1.88017 x - 0.321763 x^2 + 0.0357721 x^3

CONCLUSIONS

A new Mathematica based computer program is developed and presented as possible "shareware" for use in building by others a general purpose data-fitting model. It is envisioned that this program will develop over time as we and others build new families of basis functions into the program as well as inserting difference error measure minimization techniques as well. We hope that practitioners and students find this "shareware" type of program useful and will invest their time to contribute to the program. New code will be included on our web-site, under construction, with the contributors' names attached. The potential for improvement is quite substantial. The list of basis functions could grow dramatically. More goodness of fit measures could be added. Statistical inference of each parameter estimate could be added. We intend with this project to get the process started so that the community at large could add to it in a 'freeware' manner as time permits.

REFERENCES

Gallant, A. Ronald, *Nonlinear Statistical Models*, John Wiley and Sons, New York, 1987.

(1) Instructor, Department of Mathematical Sciences, United States Military Academy, West Point, New York
(2) Professor, Department of Mathematical Sciences, United States Military Academy, West Point, New York
(3) Academy Professor,Department of Mathematical Sciences, United States Military Academy, West Point, New York