Final Technical Report for NASA Grant NAGW-1917

Advanced Statistical Methods for Improved Data Analysis of NASA Astrophysics Missions

Eric D. Feigelson, Principal Investigator
Department of Astronomy & Astrophysics
Pennsylvania State University
University Park PA 16802

Covering February 1 1990 through January 31 1992

Under this grant, the investigators have pursued a variety of avenues for improving the statistical analysis of astronomical data. Directions include: promulgation of existing statistical techniques, development of new techniques, and production and distribution of specialized statistical software to the astronomical community. We summarize these here, referring to the Bibliography below; the abstract of each paper is appended to this report.

Linear Regression. Linear regression is a statistical technique very commonly used in astronomy, and is crucial to research associated with the cosmic distance scale (e.g., distances to galaxies, Hubble's constant, galaxy streaming, expansion age of the universe). Though apparently simple, astronomers frequently use different methods interchangeably and incorrectly. We therefore have engaged in a significant effort to inform astronomers of the intracacies of known linear regression methods, calculate some extensions to existing methods (mainly to treat cases where the X variable is random rather than fixed), and locate existing or provide new software for all methods discussed. The extensions concern the correct propagation of regression coefficient errors for any of six ordinary least squares lines, when the regression line of a calibration sample is applied to a new sample. We also discuss a wide variety of linear regression methods for data subject to measurement errors, and for flux-limited (truncated and censored) data. The work appears in three refereed papers (Linear Regression in Astronomy I and II, the first was completed prior to this grant, plus a simulation study for small sample problems) and three short Fortran codes (SIXLIN, SLOPES and CALIB). The latter are made available to the community by email request to CODE@STAT.PSU.EDU and through the Center for Excellence in Space Data and Information Sciences at Goddard Space Flight Center.

Survival Analysis. Our work on survival analysis under this grant has been the improvement, enlargement and distribution of our large ASURV (Astronomy Survival Analysis) code. The principal changes in the new Rev. 1.1 are: changing the Kaplan-Meier maximum-likelihood extimator to that it moves in the proper direction for both upper limits and lower limits data; adding a differential or binned, Kaplan-Meier estimator; substituting hypergeometric for permutation variances in some two-sample tests, which are more 'robust' against differences in the censoring patterns; removing the Cox-Mantel two-sample test and adding the Peto-Prentice test; calculating bootstrap error estimates for the slope and intercept in Schmitt's binned linear regression method; adding a new measure of bivariate correlation for doubly-censored data, based on a generalized Spearman's rho procedure developed by co-investigator Dr. M. Akritas; streamlining the screen-keyboard interface and clarifying the printed outputs; reorganizing the Users Manual so that material not actually needed to operate the program are located in Appendices; and improving code portability, so that it runs on a Sun SPARCstation under UNIX, a DEC VAX under VMS, a personal computer under MS-DOS using Microsoft FORTRAN, an IBM mainframe under VM/CMS, and (with minor format changes) a Macintosh under MacOS. ASURV Rev. 1.1 is now being distributed from CODE@STAT.PSU.EDU. The ASURV Rev. 1.1 package was presented at the *1st Annual Conference on Astronomical Data Analysis Software and Systems (ADASS)* in November 1991 in Tucson AZ. It was also presented in a broader context of censored data

in astronomy by the PI in a lecture at the Penn State conference *Statistical Challenges in Modern Astronomy.*

New Statistical Investigations. In addition to regression methods discussed above, co-investigator Dr. Babu has completed two studies relevant to astronomical problems. One is an analysis of the limitations of bootstrap methods for different purposes; he evaluates the validity of estimating variability of means, standard deviations, slopes, etc. by applying the bootstrap to subsamples of the sample of interest. The other is the derivation of optimal nonparametric survival functions (i.e. in astronomy, luminosity functions) from censored (i.e. flux-limited) samples that contain two different populations (e.g. Seyfert I and II galaxies).

Interdisciplinary Activities. The PI and co-investigator Dr. Babu were organizers of an international conference *Statistical Challenges in Modern Astronomy* held at Penn State University in August 1991. At the conference, about 80 astronomers (including representatives of GRO, COBE, HST, NASA HQ, GSFC, Ames and other NASA projects) discussed methodological issues with about 50 statisticians (including Department chairs of Yale, Stanford, Berkeley, Michigan, and Oxford). While the conference itself was funded from other NASA and NSF grants, the grant provided salary support during the organization of the conference and editing of the proceedings. The PI and Dr. Babu were also invited to give a talk at the *ADASS* conference in November 1991 on improving the statistical treatment of astronomical data.

Publications and Software Produced under Grant

Refereed Articles

"Analytical and Monte Carlo Comparisons of Six Different Linear Least Squares Fits", Gutti Jogesh Babu and Eric D. Feigelson, *Communications in Stat., Simulation and Computation* in press (1992).

"Linear Regression in Astronomy. II.", Eric D. Feigelson and Gutti Jogesh Babu, *Astrophys. J.*, submitted (Dec. 1991).

"Public Domain Software for the Astronomer: An Overview", Eric D. Feigelson and Fionn Murtagh, *Publ. Astro. Soc. Pacific*, submitted (Dec. 1991)

"Nonparametric Estimation of Survival Functions under Dependent Competing Risks", M. Bhaskara Rao, G. Jogesh Babu and C. Radharkrishna Rao, *Nonparametric Functional Estimation and Related Topics*, G. Roussas (ed.), Dordrecht:Kluwer, 1991.

"Subsample Methods", Gutti Jogesh Babu, submitted for publication (1991).

Non-refereed Articles

"Censored Data in Astronomy Due to Nondetections", Eric D. Feigelson, in *Statistical Challenges in Modern Astronomy* (eds. E. D. Feigelson and G. J. Babu), Springer-Verlag, in press (1992).

"ASURV: Astronomy Survival Analysis Package", M. LaValley, T. Isobe, and Eric Feigelson, in *Data Analysis Software and Systems* (eds. D. Worrall et al.), Pub. As. Soc. Pacific Conf., in press (1992).

"A Short Review of Sources of Public Domain Software", Fionn Murtagh and Eric D. Feigelson, in *Data Analysis Software and Systems* (eds. D. Worrall et al.), Pub. As. Soc. Pacific Conf., in press (1992).

"Improving the Statistical Methodology of Astronomical Data Analysis", Eric D. Feigelson and Gutti Jogesh Babu, in *Data Analysis Software and Systems* (eds. D. Worrall et al.), Pub. As. Soc. Pacific Conf., in press (1992).

Software Distributed

SIXLIN, 300-line Fortran program providing regression coefficients and error analysis for six different least-squares lines. Based on Isobe et al. (1990). Distributed upon request to ~110 groups in ~18 countries between June 1990 and January 1992.

SLOPES, 650-line Fortran program, extending SIXLIN to include bootstrap and jackknife error analysis for small samples. Based on Babu and Feigelson (1992). Distribution starting mid-1992.

CALIB, 500-line Fortran program, applying generalized Working-Hotelling confidence intervals to linear regression calibration problems where the X variable is random. Based on Feigelson and Babu (1992). Distribution starting mid-1992.

ASURV, Rev. 1.1, 15,000 line Fortran program improving and extending ASRUV Rev. 0. It provides a wide variety of univariate and bivariate survival analysis statistical functions treating censored data (e.g., astronomical datasets with nondetections). Distributed upon request to ~100 groups worldwide. Announced in B.A.A.S. Software Reports (1990 and 1992), and described in LaValley et al. (1992).

ANALYTICAL AND MONTE CARLO COMPARISIONS OF SIX DIFFERENT LINEAR LEAST SQUARES FITS

Gutti Jogesh Babu

Department of Statistics
219 Pond Laboratory
Pennsylvania State University
University Park PA 16802

Eric D. Feigelson

Dept. of Astronomy & Astrophysics
525 Davey Laboratory
Pennsylvania State University
University Park PA 16802

Keywords and Phrases: Tully-Fisher relation; orthogonal regression; reduced major axis; linear regression; variance estimation; cosmic distance scale.

ABSTRACT

For many applications, particularly in allometry and astronomy, only a set of correlated data points (x_i, y_i) is available to fit a line. The underlying joint distribution is unknown, and it is not clear which variable is 'dependent' and which is 'independent'. In such cases, the goal is an intrinsic functional relationship between the variables rather than E(Y|X), and the choice of least-squares line is ambiguous. Astronomers and biometricians have used as many as six different linear regression methods for this situation: the two ordinary least-squares (OLS) lines, Pearson's orthogonal regression, the OLS-bisector, the reduced major axis and the OLS-mean. The latter four methods treat the X and Y variables symmetrically. Series of simulations are described which compared the accuracy of regression estimators and their asymptotic variances for all six procedures. General relations between the regression slopes are also obtained. Among the symmetrical methods, the angular bisector of the OLS lines demonstrates the best performance. This line is used by astronomers and might be adopted for similar problems in biometry.

Linear Regression in Astronomy. II.

Eric D. Feigelson[1] and Gutti Jogesh Babu[2]

Abstract

A wide variety of least-squares linear regression procedures used in observational astronomy, particularly investigations of the cosmic distance scale, are presented and discussed. We emphasize that different regression procedures represent intrinsically different functionalities of the dataset under consideration, and should be used only under specific conditions. Discussion is restricted to least-squares approaches, and for most methods computer codes are located or provided. The classes of linear models considered are: (i) unweighted regression lines, some discussed earlier in Paper I, with bootstrap and jackknife resampling; (ii) regression solutions when measurement error, in one or both variables, dominates the scatter; (iii) methods to apply a calibration line to new data; (iv) truncated regression models, which apply to flux-limited datasets; and (iv) censored regression models, which apply when non-detections are present.

For the calibration problem, we develop two new procedures: a formula for the intercept offset between two parallel datasets, which propagates slope errors from one regression to the other; and a generalization of the Working-Hotelling confidence bands to nonstandard least-squares lines. They can provide improved error analysis for Faber-Jackson, Tully-Fisher and similar cosmic distance scale relations. We apply them to a recent published dataset, showing that the distance ratio between the Coma and Virgo clusters can be determined to $\sim 1\%$ accuracy.

The paper concludes with suggested strategies for the astronomer in dealing with linear regression problems. Precise formulation of the scientific question and scrutiny of the sources of scatter are crucial for optimal statistical treatment.

Public Domain Software for the Astronomer: An Overview

Eric D. Feigelson

Dept. of Astronomy and Astrophysics, Pennsylvania State University,
525 Davey Laboratory, University Park PA 16802, USA

Fionn Murtagh[1]

ST-ECF, European Southern Observatory, Karl-Schwarzschild-Str. 2,
D-8046 Garching, Germany

Abstract

We describe sources of public domain (PD) software available over research wide-area networks, journals and government sources which may be valuable to the astronomer and astrophysicist. A very large amount of high quality PD software is accessible at all times. We concentrate on locations with material useful for research, and offer practical suggestions regarding access in an Appendix.

Key words: Data-Handling Techniques; General Notes; Miscellaneous

Gutti Jogesh Babu
Department of Statistics
219 Pond Laboratory
The Pennsylvania State University
University Park, PA 16802

ABSTRACT

Hartigan's subsample and half-sample methods are both shown to be inefficient methods of estimating the sampling distributions. In the sample mean case the bootstrap is known to correct for skewness. But irrespective of the population, the estimates based on subsamples method and half-sample methods, have skewness factor zero. This problem persists even if we take only samples of size less than or equal to half of the original sample. For linear statistics it is possible to correct this by considering estimates based on subsamples of size λn, when the sample size is n. In the sample mean case λ can be taken as $0.5(1-1/\sqrt{5})$. In spite of these negative results, half-sample method is useful in estimating the variance of sample quantiles. It is shown that this method gives as good an estimate as that given by the bootstrap method. At the same time, half sample method is computationally more efficient. It requires less than $0(2^n n^{-1/2})$ computations, and bootstrap requires about $0(n^n)$ computations.

A major advantage of half-sample method is that it is shown to be robust in estimating the mean square error of estimators of parameters of a linear regression model when the errors are heterogeneous. Bootstrap is known to give inconsistent results in this case; although, it is more efficient in the case of homogeneous errors.

NONPARAMETRIC ESTIMATION OF SURVIVAL FUNCTIONS UNDER DEPENDENT COMPETING RISKS

M. BHASKARA RAO
Department of Statistics
North Dakota State University
FARGO, ND 58105, USA

G. JOGESH BABU* and C. RADHAKRISHNA RAO
Department of Statistics
Pennsylvania State University
UNIVERSITY PARK, PA 16802, USA

ABSTRACT

In a certain target population, the individuals will die due to either Cause 1 or Cause 2 with probabilities π and $(1-\pi)$, respectively. Let F_1 and F_2 be the life time distributions of individuals who die off due to Causes 1 and 2, respectively. In any random sample of individuals from the population, subjects can leave the study at random times. In this paper, we derive nonparametric estimates of π, F_1 and F_2 using such censored data and study some their properties. The model that suggests itself encapsulating the essentials of the problem is more general than the usual competing risks model.

10

Censoring in Astronomical Data Due to Nondetections

Eric D. Feigelson[1]

ABSTRACT Astronomical surveys often involve observations of pre-selected samples of stars or galaxies at new wavebands. Due to limited sensitivities, some objects may be undetected leading to upper limits in their derived luminosities. Statistically, these are left-censored data points. We review the nature of this problem in astronomy, the successes and limitations of using established 'survival analysis' univariate and bivariate statistical techniques, and discuss the need for further methodological development. In particular, astronomical censored datasets are often subject to experimentally known measurement errors (which are used to set censoring levels), may suffer simultaneous censoring in several variables, may have particular 'quasi-random' censoring patterns and parametric distributions.

10.1 Introduction

10.1.1 Origin of Astronomical Censoring

Consider the following situation: an astronomer goes to a telescope to measure a certain property of a preselected sample of objects. The scientific goals of the experiment might include finding the luminosity function of the objects, comparing this luminosity function to that of another sample, relating the measured property to other previously known properties, quantification of any relation by fitting a straight line, and comparing the measured property to astrophysical theory. In the parlance of statistics, the astronomer needs to estimate the empirical distribution function, perform two-sample tests, correlation and regression, and goodness-of-fit tests. Most astronomers are familiar with simple statistical methods (e.g. [Be69, Pr86]) to perform these tasks. However, these standard methods are not applicable when some of the targetted objects are not detected. In this case, the astronomer does not learn the value of the property, but rather that the value is LESS than a certain level corresponding to the sensitivity of the

[1]Department of Astronomy & Astrophysics, Pennsylvania State University, University Park PA 16802. Email: edf@astro.psu.edu

ASURV: ASTRONOMY SURVIVAL ANALYSIS PACKAGE

M. LAVALLEY
Statistics Department, Penn State University, University Park PA 16802

T. ISOBE
Center for Space Research, Massachusetts Institute of Technology, Cambridge MA 02139

ERIC FEIGELSON
Department of Astronomy and Astrophysics, Penn State University, University Park PA 16802

BACKGROUND

Observational astronomers frequently encounter the situation where they observe a particular property (*e.g.* far-IR emission in spiral galaxies, X-ray emission in young stars, CO emission in starburst galaxies) of a previously defined sample of objects, but fail to detect all of the objects. The data set then contains nondetections as well as detections, preventing the use of simple and familiar statistical techniques in the analysis.

A number of astronomers have recently recognized the existence of statistical methods, or have derived similar methods, to deal with these problems. The methods are collectively called 'survival analysis' and nondetections are called 'censored' data points. These methods recover important information implicit in the failure to detect some objects under reasonable mathematical assumptions. ASURV is a menu-driven stand-alone computer package designed to assist astronomers in using methods from survival analysis. Rev. 1.0 of ASURV provides all of the functions described in Schmitt (1985), Feigelson and Nelson (1985) and Isobe, Feigelson, and Nelson (1986), plus some additional calculations.

METHODS AVAILABLE IN ASURV

The statistical methods for dealing with censored data might be divided into a 2x2 grid: parametric *vs.* nonparametric, and univariate *vs.* bivariate. We have chosen to concentrate on nonparametric models, since the underlying distribution of astronomical populations is usually unknown.

Univariate Methods

The Kaplan-Meier estimator gives the distribution function of a randomly censored sample. First derived in 1958, it is the unique, self-consistent, generalized maximum-likelihood estimator for the population from which the sample was drawn. In its cumulative form it has analytic asymptotic (for large N) error

A SHORT REVIEW OF SOURCES OF PUBLIC DOMAIN SOFTWARE

FIONN MURTAGH[1]
Space Telescope – European Coordinating Facility, ESO, Karl-Schwarzschild-Str. 2, D-8046 Garching, Germany

ERIC D. FEIGELSON
Dept. of Astronomy & Astrophysics, Penn State University, University Park PA 16802 USA

INTRODUCTION

Software production for observational astronomy has become more efficient during the last decade with the production of large centralized systems like IRAF, MIDAS or AIPS. However, virtually all of the code is produced by astronomers, with few algorithms and little software obtained from outside the astronomical community. This wastes the limited skilled labor resources of astronomers (see Voigt and Smith 1989), and unnecessarily restricts data analysis to familiar methods. A major reason for the failure to use preexisting methods developed for other applications is the difficulty in locating the relevant software. There is no central clearinghouse for scientific software. In a very preliminary effort, we provide here an overview of some of the software resources available to astronomers from "public domain" (PD) sources. PD here means that the source code is available for scholarly use without restriction, and is either free or provided at very low cost (usually to defray distribution expenses). Our overview is restricted to software available from wide-area networks like Internet, from government repositories, and from journals. In particular, we omit the substantial body of PD software associated with scholarly monographs, bulletin boards, and we omit commercial software.

ON-LINE OR NETWORK SOURCES

Here we outline software depositories, and related on-line services such as software discussion groups, available on wide-area networks.

Netlib: A primary network source of high- and medium-quality numerical analysis software (see Dongarra and Grosse 1987 for a description). The one-line command **send index**, sent to *netlib@research.att.com*, gives sufficient information to bootstrap oneself.

Statlib: The same one-line bootstrapping message can be sent to *statlib@temper.stat.cmu.edu*, which houses extensive statistical algorithms, including many from the journal *Applied Statistics* and many in the *S* language.

---

[1]Affiliated to Astrophysics Division, Space Science Department, European Space Agency.

IMPROVING THE STATISTICAL METHODOLOGY OF ASTRONOMICAL DATA ANALYSIS

ERIC D. FEIGELSON
Dept. of Astronomy & Astrophysics, Penn State University, University Park PA 16802

GUTTI JOGESH BABU
Dept. of Statistics, Penn State University, University Park PA 16802

ABSTRACT Contemporary observational astronomers are generally unfamiliar with the extensive advances made in mathematical and applied statistics during the past several decades. Astronomical problems can often be addressed by methods developed in statistical fields such as spatial point processes, density estimation, Bayesian statistics, and sampling theory. The common problem of bivariate linear regression illustrates the need for sophisticated methods. Astronomical problems often require combinations of ordinary least-squares lines, double-weighted and errors-in-variables models, censored and truncated regressions, each with its own error analysis procedure. The recent conference *Statistical Challenges in Modern Astronomy* highlighted issues of mutual interest to statisticians and astronomers including clustering of point processes and time series analysis. We conclude with advice on how the astronomical community can advance its statistical methodology with improvements in education of astrophysicists, collaboration and consultation with professional statisticians, and acquisition of new software.

ADVANCED STATISTICS AND OBSERVATIONAL ASTRONOMY

Modern physical scientists have had little exposure to, and typically express little interest in, statistical methodology. This may be due to an underlying approach to empirical science in which the experiment is perfected until the results can be convincingly demonstrated without much statistical treatment of the data. A number of valuable books designed to help physical scientists with statistical analysis are available (e.g. Bevington 1969; Martin 1971; Eadie *et al.* 1971; Box *et al.* 1978; Press *et al.* 1986). But they generally cover only a limited range of methods and are not all widely read. Observational astronomers are thus typically exposed only to a few simple methods during their training.

Astronomical data analysis and interpretation often require quite complex and specialized statistical techniques which are not usually associated with physical experimentation. The objects astronomers observe can not be manipulated (as a chemist might purify a compound), and the experimental conditions can

ASURV Rev. 1.0

Astronomy SURVival Analysis

A Software Package for Statistical Analysis of Astronomical Data Containing Nondetections

Takashi Isobe (Center for Space Research, MIT)
Michael LaValley (Dept. of Statistics, Penn State)
Eric Feigelson (Dept. of Astronomy & Ap., Penn State)

Available from:

code@stat.psu.edu
or
Eric Feigelson
Dept. of Astronomy & Astrophysics
Pennsylvania State University
University Park PA 16802
(814) 865-0162
Email: edf@astro.psu.edu (Internet)

TABLE OF CONTENTS