The Collaborative and Methodologic Research

of

Dr. Stewart Anderson
Associate Professor of Biostatistics
University of Pittsburgh
Graduate School of Public Health

18 February 2005

Dissertation Problems

- Problem #1: How can one characterize two correlated processes which are measured simultaneously? How can one also find the instanteous rates of change for those processes with derivatives with the purpose of finding the greatest rate of change?
- Problem #2: Is there a way that flexible, semi-parametric models can be incorporated into the random effects for modeling longitudinal data?

Measurements of Oxygen Uptake in an Exercising Subject

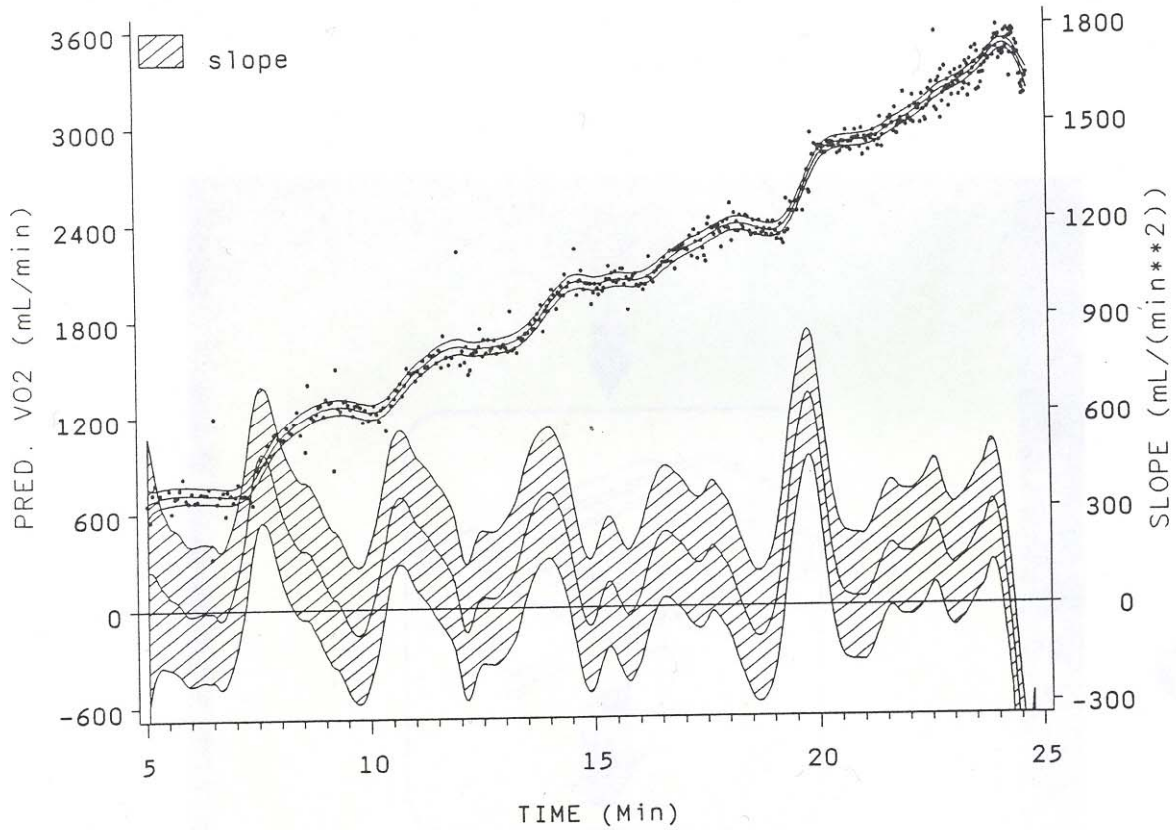


FIG. 1. *Carbon dioxide production as a function of time under increasing work for subject* B.D. *as modeled by a cubic smoothing spline. The lower curve with the wider confidence interval shows the estimated slope.*

Measurements of CO_2 Production in an Exercising Subject

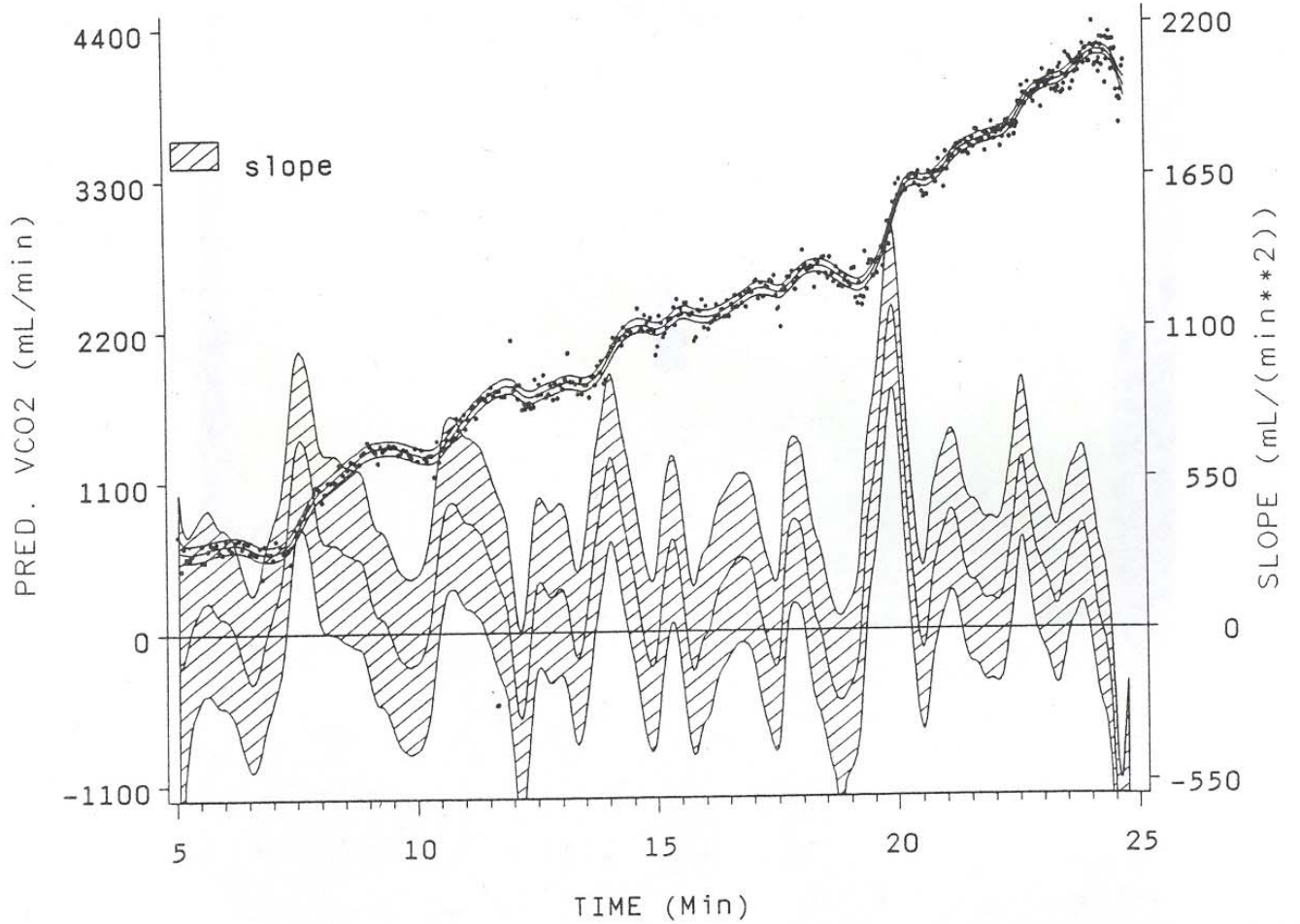


FIG. 2. *Oxygen consumption as a function of time under increasing work for subject* B.D. *as modeled by a cubic smoothing spline. The lower curve with the wider confidence interval shows the estimated slope.*

Vector smoothing splines

SIAM J. SCI. STAT. COMPUT.
Vol. 11, No. 4, pp. 749–766, July 1990

010

SMOOTHING POLYNOMIAL SPLINES FOR BIVARIATE DATA*

STEWART J. ANDERSON†, RICHARD H. JONES‡, AND GEORGE D. SWANSON§

Abstract. An extension of the smoothing polynomial spline to fit bivariate response data is presented. The data are modeled as integrated random walks with observational errors. Correlation can exist in the random walks, the observational errors, or both. The Kalman filter is used to calculate the log likelihood of the data as a function of the unknown parameters in the covariance matrices, and nonlinear optimization is used to obtain maximum likelihood estimates of the parameters. A modification of the Kalman filter is used at the beginning of the data to allow the use of diffuse (noninformative) priors. This model is applied to the problem of characterizing gas exchange time series of exercising subjects.

Key words. smoothing polynomial splines, vector splines, bivariate response models, Kalman filter, maximum likelihood estimation, integrated random walks, gas exchange measurements, Fieller's theorem

AMS(MOS) subject classification. 62-07

Using smoothing splines in a longitudinal model

STATISTICS IN MEDICINE, VOL. 14, 1235–1248 (1995)

SMOOTHING SPLINES FOR LONGITUDINAL DATA

STEWART J. ANDERSON

Department of Biostatistics, Graduate School of Public Health, 302 Parran Hall, University of Pittsburgh, 130 Desoto Street, Pittsburgh, Pennsylvania 15261, U.S.A.

AND

RICHARD H. JONES

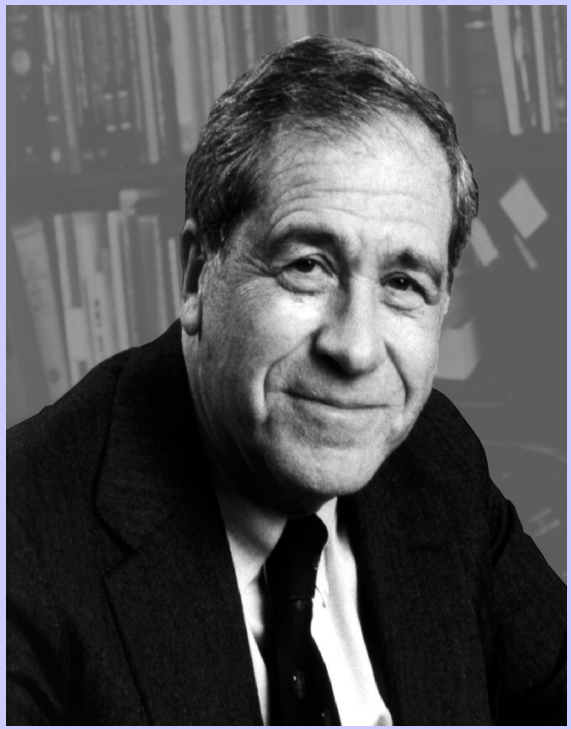
Department of Preventive Medicine and Biometrics, School of Medicine, Box B-119, University of Colorado Health Sciences Center, 4200 E. Ninth Avenue, Denver, Colorado 80262, U.S.A.

SUMMARY

In a longitudinal data model with fixed and random effects, polynomials are used to model the fixed effects and smoothing polynomial splines are used to model the within-subject random effect curves. The splines are generated by modelling the data for each subject as observations of an integrated random walk with observational error. The initial conditions for each subject's deviation from the fixed effect curve are assumed to have zero mean and arbitrary covariance matrix which is estimated by maximum likelihood, producing an empirical Bayes estimate. This is in contrast to modelling a single curve using a diffuse prior. An example is presented using unbalanced longitudinal data from a pilot study in breast cancer patients.

In 1989, I became a faculty member in the Dept of Biostatistics at the University of Pittsburgh and also became a part of

NSABP



Some Collaborations through the NSABP

- Fisher, B., **Anderson, S.**, Fisher, E., Redmond, C., et al. The significance of breast tumor recurrence following lumpectomy for the treatment of breast cancer: findings from NSABP B-06. ***The Lancet***, **338**, 327-331, 1991.
- Fisher, B., **Anderson, S.**, Redmond, C., Wolmark, N., Wickerham, D.L., and Cronin, W. Reanalysis and results after twelve years of follow-up in of a randomized clinical trial comparing total mastectomy to lumpectomy with and without irradiation in the treatment of breast cancer. ***The New England Journal of Medicine***, **333**, 1456-1461, 1995.
- Fisher, B., Jeong, J.H., **Anderson, S.**, Bryant, J., Fisher, E.R. and Wolmark, N. Twenty-five-year follow-up of a randomized trial comparing radical mastectomy, and total mastectomy followed by irradiation. ***The New England Journal of Medicine***, **347(8),** 567-575, 2002.
- Fisher, B., **Anderson, S.**, Bryant, J., Margolese, R.G., Deutsch, J., Fisher, E.R., Jeong, J.H. and Wolmark, N. Twenty-year follow-up of a randomized trial comparing total mastectomy, lumpectomy, and lumpectomy followed by irradiation for the treatment of invasive breast cancer. ***The New England Journal of Medicine***, **347(16),** 1233-1241, 2002.
- Bear, HD, **Anderson, S**, Brown, A, Smith R, Mamounas, EP, Fisher B,Margolese, R, Theoret, H, Soran, A,Wickerham, DL and Wolmark, N The Effect on Tumor Response of Adding Sequential Preoperative Docetaxel (Taxotere) to Preoperative Doxorubicin and Cyclophosphamide (AC): Preliminary Results from National Surgical Adjuvant Breast and Bowel Project (NSABP) Protocol B-27. ***Journal of Clinical Oncology***, **21(22),** 4165-4174, 2003.
- Fisher, B, Jeong JH, Bryant, J, **Anderson, S**, Dignam, J and Wolmark, N. Treatment of lymph node-negative, estrogen receptor-positive breast cancer: long-term findings from National Surgical Adjuvant Breast and Bowel Project clinical trials. ***The Lancet***, **364,** 858-868, 2004.
- Fisher ER, **Anderson S**, Dean S, Dabbs D, Fisher B, Siderits, R, Pritchard, J, Pereira, T, Geyer, C, and Wolmark, N. Solving the dilemma of the immunohistochemical and other methods used for scoring ER and PR receptors in patients with invasive breast cancer. ***Cancer***, 164-173, January 1, 2005.

Statistical Research

with Students

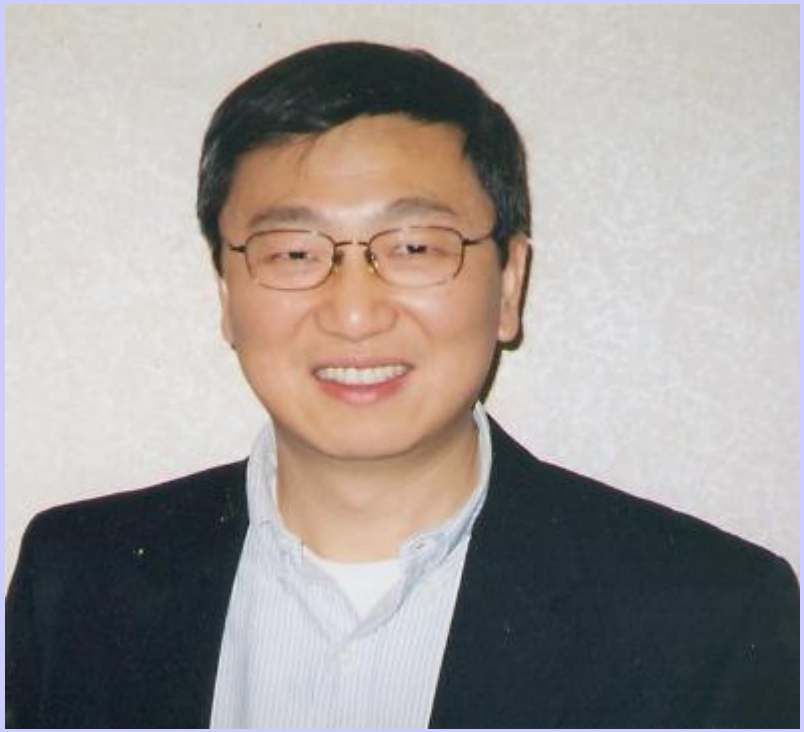
A Problem in the Analysis of Longitudinal Data

- How can one model repeated continuous (longitudinal) outcomes when
  1) more than one outcome is measured each time point? ; and
  2) the outcomes are not necessarily measured at equal time intervals?

Lingshi Tan, Ph.D.

Graduated April 1993

Dissertation: "A multivariate growth curve model with random effects and CAR(1) errors"



To appear in *Communications in Statistics*, 2005

MODELING UNEQUALLY SPACED BIVARIATE GROWTH CURVE DATA USING A KALMAN FILTER APPROACH

Qianyu Dang[1], Stewart Anderson[2], Lingshi Tan[3] and Sati Mazumdar[2]

[1]Center for Research on Health Care School of Medicine,University of Pittsburgh
Pittsburgh, Pennsylvania 15213 dangq@upmc.edu
[2]Department of Biostatistics Graduate School of Public Health,University of Pittsburgh
Pittsburgh, Pennsylvania 15261
[3]Pfizer, Inc, New York, New York 10017, U.S.A.

Key Words: Kalman filter, State space approach, Longitudinal data, Growth curve, Multivariate mixed effects model.

ABSTRACT

In many clinical studies, patients are followed over time with their responses measured longitudinally. Using mixed model theory, one can characterize these data using a wide array of across subject models. A state-space representation of the mixed e_ects model and use of the Kalman filter allows one to have great flexibility in choosing the within error correlation structure even in the presence of missing or unequally spaced observations. Furthermore, using the state-space approach, one can avoid inverting large matrices resulting in efficient computation. The approach also allows one to make detailed inference about the error correlation structure. We consider a bivariate situation where the longitudinal responses are unequally spaced and assume that the within subject errors follows a continuous first order autoregressive (CAR(1)) structure. Since a large number of nonlinear parameters need to be estimated, the modeling strategy and numerical techniques are critical in the process. We developed both a Visual Fortran and a SAS program for modeling such data. A simulation study was conducted to investigate the robustness of the model assumptions. We also use data from a psychiatric study to demonstrate our model fitting procedure.

A Sample Size Problem

- In survival analysis, how can one estimate sample size and power when a study has more than two arms and when the study has staggered accrual, drop-out, drop-in and non-proportional hazards?

Sang Ahnn, Ph.D.

Graduated December 1994.

Dissertation: “Sample size determination for comparing more than two survival distributions”

STATISTICS IN MEDICINE
Statist. Med. 17, 2525–2534 (1998)

SAMPLE SIZE DETERMINATION IN COMPLEX CLINICAL TRIALS COMPARING MORE THAN TWO GROUPS FOR SURVIVAL ENDPOINTS

SANG AHNN[1] AND STEWART J. ANDERSON[2*]

[1]*Department of Preventive Medicine, School of Medicine, State University of New York at Stony Brook, Stony Brook, New York 11794-8036, U.S.A.*

[2]*Department of Biostatistics, Graduate School of Public Health, 302 Parran Hall, University of Pittsburgh, 130 DeSoto Street, Pittsburgh, Pennsylvania, 15261, U.S.A.*

SUMMARY

This paper presents a sample size formula for testing the equality of k (≥ 2) survival distributions using the Tarone–Ware class of test statistics in the presence of non-proportional hazards, time dependent losses, non-compliance and drop-in. This method extends the derivation by Lakatos of a sample size formula for comparing two survival distributions. A sample size formula is also presented for the stratified logrank test. We describe how one can utilize these generalized formulae in calculating sample sizes and assessing power in complex multi-arm clinical trials.

Problem in assessing toxicity

- How can one effectively characterize multiple interrelated toxicities over time?

Wei Tian, Ph.D.

Graduated June 1999.

Dissertation: “Markov chain models for analyzing multivariate repeated categorical data with incomplete observations”

Published in *Communications in Statistics, Simulation and Computation*, **29(4),** 2000.

MARKOV CHAIN MODELS FOR MULTIVARIATE REPEATED BINARY DATA ANALYSIS

Wei Tian[1] and Stewart J. Anderson[2]

[1]Quintiles, Inc., P.O. Box 13979, Research Triangle Park, NC 27709

[2]Department of Biostatistics, University of Pittsburgh
Pittsburgh, PA 15261

Key Words: Markov chain models, Multivariate repeated data, Loglinear models, Clinical trials, Toxicity.

ABSTRACT

Repeated categorical outcomes frequently occur in clinical trials. Muenz and Rubinstein (1985) presented Markov chain models to analyze binary repeated data in a breast cancer study. We extend their method to the setting when more than one repeated outcome variable is of interest. In a randomized clinical trial of breast cancer, we investigate the dependency of toxicities on predictor variables and the relationship among multiple toxic effects.

A problem concerning classification trees in survival analysis

- How can one incorporate interval censoring into tree-structured models for the analysis of survival data?
  - Does interval censoring affect tree-structured models when the underlying failure rate is exponential?
  - How can one incorporate interval censoring into a nonparametric tree-structured survival model.

Yanming Yin, Ph.D.

Graduated April 2002

Dissertation: “Tree-structured modeling for interval-censored survival data”

American Statistical Association Proceedings of the Biopharmaceutical Section,
Atlanta, GA. August, 2001.

Exponential Tree-Structured Modeling for Interval-Censored Survival Data

Yanming Yin and Stewart J. Anderson, University of Pittsburgh
303 Parran Hall, GSPH, 130 DeSoto Street, Department of Biostatistics
University of Pittsburgh, Pittsburgh, PA 15261 (yayst5@pitt.edu and sja@pitt.edu)

Abstract

A recursive partitioning algorithm is proposed for interval-censored time to event data. Assuming an underlying exponential failure distribution, we use the log-likelihood as a splitting criterion. Our method is an extension of the method proposed by R. Davis and J. Anderson (Statistics in Medicine, 947-961, 1989). The performance of our method is evaluated through simulation. Specifically, we examine how sample size, proportion of right censoring and length of interval affect the performance of the partitioning. In addition, an example is given to illustrate our method.

A problem in group sequential monitoring of a clinical trial

- How can one formally incorporate both efficacy and safety outcomes into the monitoring strategy of a clinical trial?
- How should one construct "crossing boundaries" for the interim monitoring of a trial comparing an experimental treatment which is superior in efficacy but inferior in toxic side effects as compared to the standard treatment?

Maria K. Mor, Ph.D.

Graduated April 2003

Dissertation: “A Bayesian Group Sequential Approach for Multiple Endpoints”

Sequential Analysis, **24,** 1-18, 2005

A BAYESIAN GROUP SEQUENTIAL APPROACH FOR MULTIPLE ENDPOINTS

Maria K. Mor
and
Stewart J. Anderson
Department of Biostatistics
University of Pittsburgh Graduate School of Public Health
Pittsburgh, PA 15261

ABSTRACT

Making decisions regarding overall treatment effectiveness can be problematic when one considers multiple outcomes especially if the treatment effects are discordant across the profile of outcomes. A typical example involves a case where a treatment has both increased efficacy in one endpoint, e.g., improved disease-free survival, and increased side effects, e.g., more acute toxicities. Often a study is designed to test one primary hypothesis while other outcomes vital to the decision-making process are not formally incorporated into the study design. We describe a Bayesian approach that provides a mechanism for combining information from two normally distributed endpoints and accounts for the magnitude of those effects. This procedure is implemented for the case of comparing two different treatments to each other and allows for multiple looks at the data. Information from more than one endpoint is combined through the use of utility functions. Our group sequential procedure is demonstrated for the design of a cancer clinical trial that involves two looks at the data. The example shows the effect of different utility functions applied to the same data. Because the selection of the utility function is crucial to the interpretation of the two endpoints, the results are not invariant to the utility function, and great care must be exercised in choosing an appropriate function. Additionally, since results are not robust to the choice of prior, we select a non-informative prior for our example. Despite some limitations in the specification of the utility structure and prior distribution, our procedure provides a mechanism that is useful for simultaneously monitoring multiple endpoints in a clinical trial.

KEY WORDS: Bayesian methods; group sequential designs; multiple endpoints; clinical trials

Mary Kelley, Ph.D

Graduated December 2003

Dissertation: “Zero Inflation in Ordinal Data: Applications of a Mixture Model”

ZERO INFLATION IN ORDINAL DATA: APPLICATIONS OF A MIXTURE MODEL **

Mary Elizabeth Kelley, PhD
University of Pittsburgh, 2003

The aim of the current proposal is to produce a methodology that will allow users of ordinal scale data to more accurately model the distribution of ordinal outcomes when it is assumed that not all patients will exhibit a response (i.e., exhibits zero inflation). This situation occurs with ordinal scales in which there is an anchor that represents the absence of the symptom or activity, such as "none", "never" or "normal", and is particularly common when measuring abnormal behavior, symptoms, and side effects. Due to the unusually large number of zeros, traditional statistical tests of association can be non-informative. We propose a mixture model for ordinal data with a built-in probability of non-response. Simulations show that the model is well behaved and information criterion can be used to choose between the zero-inflated and the traditional proportional odds model. The model, however, does have restrictions on the nature of the covariates that must be satisfied in order for the model to be identifiable. If the appropriate restrictions hold, the mixture model proposed allows modeling of the range (e.g., severity) of the scale, while simultaneously modeling the presence/absence of the symptom, and allows the predictors of these two aspects of the outcome to differ. This is particularly relevant to public health research methods such as large epidemiological surveys.

****Manuscript in preparation**

Michael Brent McHenry, Ph.D.

Graduated June 2004

Dissertation: “New Estimation Approaches in Survival Analysis with Aalen’s Additive Risk Model”

A problem associated with the use additive models in survival analysis

- Can one minimize problems of negative hazard estimates in Aalen's additive model by smoothing the cumulative hazard?
- How can splines be used to characterize the risk (hazard) of an event over time in an additive model?
- Is there a way of constraining Aalen's additive model in such a way that the hazard estimates are always nonnegative?

Submitted to *Biostatistics* in September 2004, Currently under revision

Estimation for Aalen's Additive Risk Model Using Smoothing Polynomial Splines

M. BRENT MCHENRY, STEWART J. ANDERSON
Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261, U.S.A.
CHUNG-CHOU H. CHANG
Department of Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, U.S.A.
LISA A. WEISSFELD
Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261, U.S.A.
MARK S. ROBERTS
Department of Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, U.S.A.

SUMMARY

A use of smoothing polynomial splines (SPS) is proposed to estimate a regression function (coefficient) from the corresponding ordinary least-squares (OLS) cumulative regression function for Aalen's additive risk model. With the SPS method, the generalized maximum likelihood can be used to directly estimate the bandwidth for a given dataset. The bias and mean squared error (MSE) of the SPS method are comparable to the bias and MSE of the Epanechnikov kernel (E-KER) method used by Aalen (1993). Both the SPS and the E-KER methods allow analysts to make inferences about the rates of change of the cumulative regression function, but only the SPS method allows analysts to make inferences about the rates of change of the regression functions.

*Keywords***:** Aalen additive risk model; Cumulative regression function; Kernel smoothing; Penalized least-squares; Regression function; Smoothing polynomial spline.

All my Ph.D. Primary Advisees (Picture taken at ENAR 2004)