



**SACRAMENTO
STATISTICAL
ASSOCIATION**

A local chapter of the American Statistical Association

**PROGRAM
Institute on Research and Statistics**

Wednesday, March 15, 2006
California State University, Sacramento

<http://www.amstat.org/chapters/sacramento/>

President: Linda Gage 327-0103 x2549; Vice-President: Chris Drake 530-752-8170 , Secretary: Kathleen Gallagher 552-9642; Treasurer: Charles Chan 552-9694; ASA Representative: Linda Gage 327-0103 x2549; Past President: Matthew Facer 449-5835; Councilors: Jennifer Baham 449-5853; Prabir Burman 530-752-7622; Shannon Conroy 449-5280; Kirsten Knutson 449-5305; Lois Lowe 722-3310; Doraiswamy Ramachandran 278-6534; Gloria Robertson 654-1837

All events will take place at the California State University Union

Registration and Continental Breakfast	Redwood Room	8:00-9:00
-----------------------------------------------	---------------------	------------------

Morning Plenary Sessions	Redwood Room	9:00-10:55
---------------------------------	---------------------	-------------------

9:00-9:15

Welcoming Remarks

Linda Gage, President, Sacramento Statistical Association
Doraiswamy Ramachandran, Department of Mathematics and Statistics, CSUS

9:15-10:05

Special Guest Speaker

Fritz Scheuren, Past-President, American Statistical Association
"Our Pro Bono Obligation"

10:05-10:55

Featured Speaker

Rosemary Cress, Research Program Director, California Cancer Registry/Public Health Institute
"The California Cancer Registry: California's Cancer Surveillance System"

I. Concurrent Sessions	11:00-11:55
-------------------------------	--------------------

California Suite (3rd floor)

(11:00-11:50) Julie Smith-Gagen, Center for Health Data and Research, Nevada State Health Division,
Investigation of high surgical volume survival advantages using relative survival to estimate rectal cancer deaths in a California population based analysis.

To investigate the association of survival among rectal cancer patients and provider characteristics, we used the California Cancer Registry to identify 5583 patients diagnosed with localized or regional rectal cancer from 1993 through 1996 who underwent surgical resection. Patients were followed through 2001, with follow-up estimated to be over 94% complete. Provider characteristics assessed included hospital medical school affiliation, self-identified surgeon specialty, and hospital and surgeon surgical volume. We calculated relative survival rates because they avoid the need for cause-of-death information, which is often missing or misclassified for rectal cancer and requires researchers to group rectal and colon cancers together despite their different risk factors and treatment regimens. We adjusted for prognostic factors with a generalized linear model based on collapsed data using the exact survival times and a Poisson assumption to utilize goodness-of-fit and regression diagnostics. Patients treated in low volume hospitals had a 30 percent relative excess risk of death compared to high volume hospitals and patients treated by low volume surgeons had a 52 percent relative excess risk compared to high volume surgeons. These effects were independent. Hospital teaching status and surgeon specialty were not significant. The risk differences were also examined. For younger patients with regional disease and low socioeconomic status (SES), those treated by low volume compared to high volume surgeons had 54 excess deaths per 1000 person-years.

Delta Suite (3rd floor)

(11:00-11:50) Raphael Diaz, Department of Statistics, University of California, Davis
Comparison of PQL and Laplace 6 Estimates of Hierarchical Generalized Linear Models when Comparing Groups of Small Incident Rates in Cluster-Randomized Trials

One of the approaches that are used to take into account the possible correlation between members in the same cluster in cluster-randomized trials with binary outcomes is that of hierarchical generalized linear models (HGLMs). The estimates of the parameters in a particular HGLM, a hierarchical logistic model, can be used to compare two groups of proportions arising from these types of trials with what is known as the logistic-normal likelihood ratio (LNLR) test. Penalized quasi-likelihood (PQL) estimation is a commonly used technique to estimate the parameters of HGLMs. However, the PQL estimates of the variance components in hierarchical logistic models have been shown to be biased. A recently proposed estimation technique, the Laplace 6 approximation, promises more accurate estimates. The PQL and the Laplace 6 estimates of the parameters in the hierarchical logistic model with which the LNLR test is conducted are compared for practical scenarios through Monte Carlo simulations, and the results of this comparison are presented in this talk. These results show that the Laplace 6 estimates of the variance components in this model are invariably less biased than those obtained with the PQL technique, but that this bias reduction is obtained at the expense of higher mean square errors. An example illustrates that these results can produce different conclusions when estimating the parameters in the LNLR test with these two techniques. Caution should be used until the properties of the Laplace 6 approximation are better established. The talk concludes with recommendations for further research in this area based on the results of this study.

Lunch Buffet	Redwood Room	12:00-1:00
---------------------	---------------------	-------------------

II. Concurrent Sessions		1:05-2:00
--------------------------------	--	------------------

Delta Suite (3rd floor)

(1:05-1:30) Yanhua Zhang, Department of Statistics at University of California at Davis.
Consistent Model and Moment Selection Criteria for GMM Estimation: An application of FENCE Method to financial Dynamic Panel Data Models

This paper proposes a consistent model and moment selection criterion for GMM (Generalized Method of Moments) estimation. This criterion is an application of FENCE Method, which constructs a statistical fence to eliminate incorrect models. This paper applies the model and moment selection criterion to a dynamic panel data from financial markets in selecting the number of moment conditions and the lag length of lagged dependent variables. The criterion puts a premium on the usage of fewer parameters for a given number of moment conditions and more moment conditions given the number of parameters. The results of a Monte Carlo experiment on a dynamic panel data model and a real-life data analysis are reported to show the finite sample performance of this model and moment selection criterion.

(1:35-2:00) Zhen Zhang, Graduate Group of Biostatistics, University of California, Davis
Density Warping

Estimation of the density function for a group (or groups) of individuals are of common interest in many fields. For example, in life-science studies, longitudinal data of a large number of cohorts are often observed. Given the densities of age-at-death of individual cohorts, a basic statistical problem would be how to come up with a reasonable underlying density. Based on empirical results from life time data, there usually exists timing-variation among cohorts. Conventional methods, such as the cross-sectional average density, often ignore this timing-shift and hence lead to an overall density that is not representative. A new proposal is to view densities as functional data where individual densities are warped into a time-warping density estimate. Our approach can be considered as an extension of the functional convex averaging method introduced by Liu and Müller (2004). In our research, we consider the observed densities as realizations of an underlying process via a warping mapping. By first targeting at the expectation of the latent process, the warped density is obtained from an inverse warping mapping. Hazard rate can be derived directly from the warped density when there was no censoring. Asymptotic properties of the warped density when using a data-driven warping function are also reported. Simulation results show clearly that the warped density overcome the drawback of conventional methods. The approach is illustrated with longevity data obtained for 143 cohorts of medflies (Mediterranean fruit flies). This is based on joint work with Prof. Hans-Georg Müller.

California Suite (3rd floor)

(1:05-1:30) Rani Celia Isaac, Franchise Tax Board, Research Bureau
Convergence in Incomes Proves Elusive

The convergence in incomes across the US has run its course and further convergence is proving to be elusive. Using a variance measure of dispersion, i.e. the standard deviation of the log of incomes, shows rapid progress between 1930 and 1979, but reversals in the trend are beginning to cause concern. The gap between rich and poor states has begun to widen again. Some factors in determining a state's relative per capita income are the presence (or absence) of particular industrial sectors, educational attainment of the labor force, and location advantages (proximity to markets or supplies). Differences in incomes feed into home values. Partially because of leverage, home values have greater differences than incomes. tba

(1:35-2:00) Yu Zhang, Columbia Business School, Department of Finance
Predictability of Expected Returns: Is It There in The Emerging Stock Markets?

The predictability of expected returns has been documented in the stock markets of developed countries. However, the literature in the emerging markets had mixed results. The financial crises in emerging markets in the 1990's complicate the research: the return returns in this period are different than those in the relatively "mild" period ending 1990. Using mixed model analysis, this paper revisits the predictability of expected returns in emerging markets in the long run, with the help of recently available data. Two factor models are presented, one of which pre-specifies the underlying factors while the other treats the factors as latent. Using Generalized Methods of Moments, the empirical testing implies that the latent-factor model performs better than the prespecified-factor model. The test of no predictability is rejected for half of the 12 emerging countries. Furthermore, contrary to what some papers have argued, the world-market portfolio doesn't seem to be a good candidate for a latent factor in pricing expected equity returns.

Keywords: Predictability of expected returns, emerging markets, mixed model, GMM, latent factors.

Special Session. An Invited Tutorial Sponsored by the SAS Institute, Inc.

1:30-3:00

Auburn Room Dr. Anthony Waclawski, SAS Institute, Inc.
Mining for Money

In corporate America a forecast is often required whenever a decision is made. This is especially true for decisions regarding management of large international portfolios. Philosophically, the ultimate purpose of describing past return on investment is to acquire baseline knowledge that will enable one to predict future performance. Data of potential value in the formulation of policy frequently occur in the form of time series. Questions of the following kind often arise: "Given a known intervention (i.e. change in the investment portfolio), is there evidence that change in the series of the kind actually expected actually occurred, and, if so what can be said of the nature and magnitude of the change?" From one perspective, one can argue that it is highly questionable that past time periods are statistically valid random sample's of all immediately succeeding future time periods. Although we may be able to precisely describe the past pattern of variation in our criterion variable there is absolutely no theoretical assurance that it will occur in the future, if at all. Fortunately, empirical investigation of time series data suggests that future events are often a reflection of historical trends. The Central Theorem holds that all physical phenomena regress towards their arithmetic mean over time and is often cited as the theoretical construct that allows one to employ stochastic variables to successfully forecast the future. Researchers have found that fundamental economic factors can be used to forecast security returns. However, what factors to include and how to model the relationship remain open questions. Financial economists have carefully selected and tested a small set of variables suggested by economic theory. At the other extreme Morillo and Pohlman (2002) forecasted equity market returns by applying the dynamic factor model of Stock and Watson (1998) to large set of macroeconomic variables. In this article we apply the latest data mining techniques to forecasting equity market returns. The results are economically significant.

Delta Suite (3rd floor)

(2:05-2:30) Alice van Ommeren, LeAnn Fong-Batkin, Research & Planning, California Community Colleges
The Use of Zip Code Level Data to Develop an Economic Service Area Index for the California Community Colleges

The economic background of enrollees at each community college is considered an important variable in analyzing the differences between the institutions. Previous research by the California Community Colleges Chancellor's Office has shown that county income is a significant predictor in the analysis of educational outcomes. The economic conditions for the actual geographic area served by a community college may differ, for various reasons, from the economic conditions for the county in which the college is located. Because the state's administrative database for the community colleges lacks a measure of the economic environment for the student population at each college, the research staff at the Chancellor's Office devised a new measure based on linking the student zip codes with 2000 Census income data. This presentation introduces the methodology of combining enrollment patterns (Fall 2000) of community college students by ZIP code of residence with income data (1999) from Zip Code Tabulation Areas (ZCTA) obtained from Census 2000 to create an economic service area index (ESAI) for each of the 109 community colleges in California. The ESAI is compared to county level income data and explored as a predictor in preliminary analysis and modeling of college level outcomes. We will discuss the benefits, as well as limitations, of using ZIP code level data versus county level data. This talk will explore the applicability of developing a similar economic index for other fields and disciplines. A list of references and resources will be provided.

(2:35-3:00) Lois Lowe, Ph.D, Consultant
Characteristics of California Adult Prisoners over 15 Years: 1989–2004

During the 1980s, the number of adult prisoners incarcerated within California Department of Corrections and Rehabilitation (CDCR) prisons increased at a rapid rate, reaching 87,297 in 1989. New commitments for substance abuse offenses contributed significantly to the increase. Further, 85% of inmates released to parole returned to custody within two years. Interventions, which included in-prison and parole substance abuse programs, were implemented beginning in 1990. Public sector data were used to determine changes in inmate characteristics over a 15-year period following the interventions. In 2004, 158,191 individuals were incarcerated within CDCR. Data by age, gender, race/ethnicity and primary commitment offense for years 1989, 1994, 1999 and 2004 show, consistent with California population changes, that the prison population has become more Hispanic/Latino and a little older. As expected, the percentage of inmates committed for a substance abuse offense has declined.

California Suite (3rd floor)

(2:05-2:30) Juan Yang,¹ Katherine E. Hartmann,^{1,3} Amy H. Herring,⁴ and David A. Savitz^{1,2}
¹California Department of Health Services, CDIC, Tobacco Control Section, Sacramento, California
²Carolina Population Center, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.
³Department of Obstetrics and Gynecology, School of Medicine, The University of North Carolina at Chapel Hill, North Carolina.
⁴Department of Biostatistics, School of Public Health, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.
Reducing Misclassification in Assignment of Timing of Events during Pregnancy

Background. Perinatal epidemiology studies often collect only the calendar month in which an event occurs in early pregnancy because it is difficult for women to recall a specific day, when queried later in pregnancy or postpartum. Lack of day information may result in incorrect assignment of completed gestational month since calendar months and pregnancy months are not aligned. **Methods.** To examine the direction and magnitude of misclassification, we compared three methods for assignment of completed gestational month: 1) calendar month difference, 2) conditional month difference, and 3) imputed month midpoint. We used data from the Pregnancy, Infection, and Nutrition Study for simulations. **Results.** Calendar month difference misclassified 54% of events as one month later in pregnancy compared with the actual completed month of gestation. Each of the other two methods misclassified about 12% of events to one month earlier and 12% to one month later. **Conclusions.** Calendar month difference, a common method, has the

greatest misclassification. Conditional month difference and imputed month midpoint, which require little effort to implement, are superior to calendar month difference for reducing misclassification.

(2:35-3:00) Xueying Zhang , Hao Tang, David Cowling, California Department of Health Services Chronic Disease & Injury Control Division/Tobacco Control Section
A Confirmatory Factor Analysis of a Social Norm Change Paradigm for California Tobacco Control Program

Objective: This study is to identify a set of latent variables underlying the attitudinal questions in the California Adult Tobacco Survey (CATS), and to develop a higher order “social norm change” paradigm that attempts to reflect the California Tobacco Control Program’s priority areas. These findings would support the theoretical foundation of social norm change and establish a measurement framework to help California monitor its comprehensive tobacco control program. **Methods:** We analyzed CATS data from 1997-2004 (n=33,907). Exploratory and confirmatory factor analyses were used to determine the nature of the underlying factors that reflect the “social norm change” paradigm and to reveal the underlying structure (latent variables) from a large set of attitudinal questions. **Results:** We found six first-order latent variables underlying the nineteen attitudinal questions. The higher order latent constructs were government regulation and protection (GRP), Second Hand Smoke (SHS), and Tobacco Industry Manipulation (TIM). The second order factor loadings showed that the final social norm change paradigm consists of three main constructs that are related to the primary activities and tools required to change social norms in tobacco control practices. **Conclusions:** The SHS and TIM constructs are core components of the California program and can be monitored using these latent variables. The GRP construct consists of various components that based on the loadings are contradictory (including product regulation, youth access and advertising restrictions). This may suggest that this area is extremely complicated to influence with a simple straightforward message since beliefs across a variety of topics are intertwined.

Afternoon Refreshments

3:00-3:15

IV. Concurrent Sessions

3:20-4:1

Delta Suite (3rd floor)

(3:20-4:15) Enoch Haga, Retired Teacher
T-Scores for Teachers, A Tool for Grading Students Fairly

Fairness in grading student achievement is of major concern to every classroom teacher at every level from primary school through the university. With the advent of commercial grading programs, teachers are expected to be always ready with tentative grades for student counseling or parental meetings. This presentation briefly explains the theory, using z- and T-scores, behind ranking students according to position on a normal curve. Because this procedure is easily explained and understood, it eliminates questions of fairness in grading students.

California Suite (3rd floor)

(3:20-4:15) Debashis Paul , Department of Statistics, University of California, Davis
Sparse principal component analysis for structured high dimensional data

Principal components analysis (PCA) is a widely used tool for reducing the dimensionality of multivariate data. Increasingly we are confronting multivariate data with very high dimension and comparatively low sample size, e.g. in medical imaging, microarray analysis, speech and image recognition, atmospheric science, finance etc. In this talk we consider the problem of estimation of the principal components in situations where the dimension of the observation vectors are comparable to the sample size, even though the intrinsic dimensionality of the signal part of the data is small. We shall demonstrate that the standard technique involving eigen-analysis of sample covariance matrix can fail to provide good estimates of the eigenvectors of the population covariance matrix. However, if the eigenvectors corresponding to the bigger eigenvalues of the population covariance matrix are sparse in a suitable sense, then one can hope to get better estimates of these components. A two-stage algorithm will be proposed that first selects a set of significant coordinates through a thresholding procedure, and then performs a PCA on the submatrix of the sample covariance matrix corresponding to the set of selected coordinates to arrive at the final estimate. The performance of the procedure will be demonstrated via some simulated and real data examples.

Networking, Socializing

4:20-5:00

The Auburn Suite is available for your use throughout the day (3rd floor) except 1:30-3:00pm