NKUA MsC in Biostatistics: Seminar on "MODERN INCOMPLETE DATA PROBLEMS"

Friday 5th May 11:00 – 14:00

Amphitheater of Physiology-G.KOTZIAS (Building 15)

Medical School of the National & Kapodistrian University of Athens, (NKUA)

75 M. Asias str, 115 27 Goudi

11:00-11:30 INVERSE PROBABILITY OF CENSORING WEIGHTS UNDER MISSING NOT AT RANDOM WITH APPLICATIONS TO LONG-TERM CLINICAL OUTCOMES

Constantin T. Yiannoutsos
Indiana University R. M. Fairbanks School of Public Health, Indianapolis, IN, USA
Email: cyiannou@iupui.edu

Background. Estimate the response to antiretroviral therapy (ART) among HIV-positive patients who start ART in sub-Saharan Africa. Dealing with death when estimating longitudinal measurements and produce a survival-adjusted longitudinal measure (e.g., median CD4 count, percent of patients with perfect ART adherence) and address counterfactual scenarios (e.g., What is the value of the measure had everyone remained on f/u but not necessarily in care).

Methods. Inverse Probability of Censoring Weighting (IPCW) methods readily available to estimate median CD4 count over time but MAR assumption likely not applicable in our setting. Use patient tracing (double-sampling) data and modify IPCW for the MNAR setting (MNAR-IPCW).

Results. Both longitudinal measures considered (median CD4 count, perfect adherence) were overestimated, even compared to the best-case scenario of everyone having remained under observation and in care.

Conclusions. Ignoring biases resulting from non-random losses to follow-up may results in significant biases when estimating longitudinal measurements. These results have broad application on a number of longitudinal biomarkers in this setting, particularly those related to the long-term viral suppression necessary to maintain good outcomes among people living with HIV around the world.

11:30-12:00 SEMIPARAMETRIC ANALYSIS OF COMPETING RISKS DATA UNDER DOUBLE-SAMPLING DESIGNS

Giorgos Bakoyannis, Ying Zhang, Constantin T. Yiannoutsos
Indiana University R. M. Fairbanks School of Public Health, Indianapolis, IN, USA
Email: gbakogia@iu.edu

Outcome diagnosis in cohort studies and clinical trials with competing risks is often expensive or invasive. In such cases, less expensive or easily applicable alternative diagnostic procedures may be used, but these procedures are usually prone to error. This can lead to outcome misclassification and, therefore, seriously biased estimates. An efficient approach to deal with the problem of expensive or invasive diagnostic procedures is double-sampling design. Under this design, an expensive gold standard failure type ascertainment procedure is only used in a small subset of non-censored observations, while a less expensive or invasive, but imperfect, alternative is being used for all the non-censored cases. For this design, we propose a computationally efficient maximum profile pseudolikelihood estimator for the semiparametric proportional cause-specific hazards model. Using modern empirical process theory we derive the asymptotic properties of the estimators for both the regression coefficients and the covariate-specific cumulative incidence functions. We provide closed-form variance estimators and also propose methodology for confidence band construction for the covariate-specific cumulative incidence function. Our methodology provides a unified approach for inference in terms of both risk factors and personalized risk predictions. Simulation studies show that the proposed estimators perform well even with small double-sampling ratio. The method is illustrated using data from an HIV study with a double-sampling design.

12:00-12:30 LONGITUDINAL AND TIME-TO-DROP-OUT JOINT MODELS CAN LEAD TO SERIOUSLY BIASED ESTIMATES WHEN THE DROP-OUT MECHANISM IS AT RANDOM

Christos Thomadakis[1,*],_Loukia Meligkotsidou[2,**], Nikos Pantazis[1,***] and Giota Touloumi[1,****]
1: Department of Hygiene and Epidemiology, University of Athens, Greece; 2: Department of mathematics, University of Athens, Greece.
Emails: *: cthomadak@med.uoa.gr; **: meligots@math.uoa.gr; ***: npantaz@med.uoa.gr; ****: gtouloum@med.uoa.gr

Missing data are common in longitudinal studies. Likelihood-based methods ignoring the missingness mechanism are unbiased provided that missingness is at random (MAR), whereas under not at random missingness (MNAR), joint modeling of the longitudinal data and the missingness mechanism is required. In our motivating example of modeling CD4 cell count trajectories during untreated HIV infection, CD4 counts are mainly censored due to treatment initiation, with the nature of this mechanism remaining debatable. In this work we evaluate the performance of a specific class of joint models termed shared-parameter models (SPMs) under MAR drop-out and propose an alternative model for modeling two correlated markers through a Bayesian MCMC procedure. We analytically calculate the asymptotic bias in two specific SPMs under specific MAR drop-out mechanisms, showing that the bias in the marker's slope increases as the MAR drop-out mechanism becomes heavier. A simulation study is carried out to evaluate the performance of the proposed model and of other commonly used SPMs, under MAR and MNAR scenarios. Under MAR, the proposed model has the best performance in terms of bias compared to the other SPMs, the estimates of which are seriously biased. Under MNAR, the estimates from the proposed model are again nearly unbiased, whereas those from the other SPMs are moderately to heavily biased, depending on the parameterization used. The examined models are also fitted to real data from the Concerted action on seroconversion to AIDS and death in Europe" (CASCADE) study. Results from the real data example are compared and discussed in the light of our analytical and simulation-based results.

12:30-13:00 DETERMING THE LIKELY PLACE OF HIV ACQUSITION FOR MIGRANTS IN EUROPE COMBINING SUBJECT-SPECIFIC INFORMATION AND BIOMARKERS DATA

Nikos Pantazis1, Christos Thomadakis1,, Julia del Amo2, Debora Alvarez-del Arco2, Fiona M Burns3,4, Ibidun Fakoya3, Giota Touloumi1, on behalf of the aMASE and CASCADE study groups

1: Department of Hygiene, Epidemiology and Medical Statistics, Medical School, National and Kapodistrian University of Athens; 2: National Centre of Epidemiology, Instituto de Salud Carlos III, Madrid, Spain; 3: Centre for Sexual Health and HIV Research, Research Department of Infection and Population Health, University College London, London, UK; 4: Royal Free London NHS Foundation Trust, London UK

E-mail: npantaz@med.uoa.gr

In most HIV-positive individuals, infection time is only known to lie between the time an individual started being at risk for HIV and diagnosis time. However, a more accurate estimate of infection time is very important in certain cases. For example, one of the objectives of the aMASE study was to determine if HIV-positive migrants, diagnosed in Europe, were infected pre- or post-migration. We

propose a method to derive subject-specific estimates of unknown infection times using information from HIV biomarkers measurements, demographic, clinical and behavioral data. We assume that CD4 cell count (CD4) and HIV-RNA viral load (VL) trends after HIV infection follow a bivariate linear mixed model. Using post-diagnosis CD4 and VL measurements and applying the Bayes rule, we derived the posterior distribution of the HIV infection time, whereas the prior distribution was informed by AIDS status at diagnosis and behavioral data. Parameters of the CD4-VL and time-to-AIDS models were estimated using data from a large study of individuals with known HIV infection times (CASCADE). Simulations showed substantial predictive ability (e.g. 84% of the infections were correctly classified as pre- or post-migration). Application to the aMASE study (n=2,009) showed that 46.0% of African migrants and 70.7% to 74.6% of migrants from other regions were most likely infected post-migration. Applying a Bayesian method based on bivariate modeling of CD4 and VL, and subject-specific information, we found that the majority of HIV-positive migrants in aMASE were most likely infected after their migration to Europe.

13:00-13:30 EVALUATING DIAGNOSTIC ACCURACY OF BIOMARKERS IN THE PRESENSE OF MISSING BIOMARKERS Shanshan Li

Indiana University R. M. Fairbanks School of Public Health, Indianapolis, IN, USA
Email: sl50@iu.edu

The Receiver Operating Characteristic (ROC) curve is a common tool for evaluating diagnostic accuracy of biomarkers. The existing work on estimating ROC functions are mostly developed for data collected under ideal settings. In many medical studies, measurements of biomarkers are subject to missingness due to high cost or limitation of technology. To deal with the missing data problem in biomarker studies, we propose an augmented weighted distribution model that incorporates information from subjects with incomplete data. The resulting estimator enjoys the double-robustness property in the sense that it remains consistent if either the missing data process or the conditional distribution of the missing data given the observed data is correctly specified. We derive the asymptotic properties of the proposed estimators and evaluate their performances using extensive numerical studies.

13:30-14:00 TWO-STAGE SEMIPARAMETRICS ANALYSIS OF SKELETAL GROWTH AROUND PUBERTAL GROWTH SPURT WITH INTERVAL-CENSORED OBSERVATIONS

Chenghao Chu, Ying Zhang and Wanzhu Tu

Indiana University R. M. Fairbanks School of Public Health, Indianapolis, IN, USA

Email: yz73@iu.edu

Human individuals acquire their adult body shapes through vigorous physical growth in the first two decades of life. Many of the somatic characteristics that define our physical appearance in adulthood take shape around the time of pubertal growth spurt (PGS). An analytical challenge to quantify growth rates before and after PGS is the lack of direct observation of the anchoring PGS event. We propose a two-stage semiparametric analysis to assess the rates of skeletal changes around the PGS with interval-censored observation on the PGS. The first stage is the nonparametric maximum likelihood estimation for the distribution of PGS timing. In the second stage, a least-squares based method is used to estimate the model parameters, including the pre and post-PGS growth rates with latent time of PGS. We show that under mild regularity conditions, the estimators are consistent and asymptotically normal. Statistical inference ensues from the large sample theory. We conduct a simulation study to evaluate the operating characteristics of the proposed method. Analysis of growth data from an observational cohort shows that in comparison to girls, boys tend to have a more sustained skeletal growth after PGS, as evidenced by the greater post-PGS growth rates in the upper body. The findings suggest that strong and sustained post-PGS skeletal growth contributes to the sexual dimorphism in human body.