# Workshop on Missing Data

**Time:** 5月22日 (Mon) ~ 23日 (Tue)

**Place:** 中央研究院統計科學研究所
台北市南港區研究院路二段128號

**主辦單位：** 中央研究院統計科學研究所

<table>
<thead>
<tr>
<th>5月22日</th>
<th>Presentation</th>
<th>Speaker</th>
<th>Place</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:30~11:30</td>
<td>Non-parametric and semiparametric models for missing covariates in parametric regression</td>
<td>陳華雲教授</td>
<td>2F交誼廳</td>
</tr>
<tr>
<td>11:30~13:00</td>
<td>午餐</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15:00~16:00</td>
<td>Expected Estimating Equations for Missing Data, Measurement Error, and Misclassification, with Application to Longitudinal Nonignorably Missing Data</td>
<td>王清雲教授</td>
<td>3F 308室</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>5月23日</th>
<th>Presentation</th>
<th>Speaker</th>
<th>Place</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:30~11:30</td>
<td>Boosting with Missing Predictors</td>
<td>王清雲教授</td>
<td>3F 308室</td>
</tr>
<tr>
<td>11:30~13:00</td>
<td>午餐</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15:00~16:00</td>
<td>Likelihood robustification in parametric regression with missing data</td>
<td>陳華雲教授</td>
<td>3F 308室</td>
</tr>
</tbody>
</table>
Non-parametric and semiparametric models for missing covariates in parametric regression

Abstracts

Robustness of covariate modeling for the missing covariate problem in parametric regression is studied under the MAR assumption. For a simple missing covariate pattern, non-parametric likelihood is proposed and is shown to yield a consistent and semiparametrically efficient estimator for the regression parameter. Total robustness is achieved in this situation. For more general missing covariate patterns, novel semiparametric models are proposed for modeling missing covariates. In this modeling approach, the covariate distribution is first decomposed into the product of a series of conditional distributions according to the overall missing data patterns and the conditional distributions are then represented in the general odds ratio form. The general odds ratios are modeled parametrically and the other components of the covariate distribution are modeled nonparametrically. Maximum semiparametric likelihood is proposed to find the parameter estimates. The proposed method yields a consistent estimator for the regression parameter when the odds ratios are modeled correctly. In general, the semiparametric covariate modeling strategy increases the robustness against covariate model misspecification when compared with the parametric modeling strategy proposed by Lipsitz and Ibrahim. The new covariate modeling approach can also be incorporated into the doubly robust procedure of Robins et al to increase protection against misspecification of the missing data mechanism. Furthermore, the proposed modeling strategy avoids the usually intractable integrations that are involved in the maximization of the incomplete data likelihood with parametric covariate models. The proposed method can be applied to solve missing covariate problems in many regression models frequently used in practice.

Expected Estimating Equations for Missing Data, Measurement Error, and Misclassification, with Application to Longitudinal Nonignorably Missing Data

Abstracts

Missing data, measurement error and misclassification are three important problems in many research fields, such as epidemiological studies. It is well known that missing data and measurement error in covariates may lead to biased estimation. Misclassification may be considered as a special type of measurement error, for categorical data. Nevertheless, we treat misclassification as a different problem from measurement error since statistical models for them are different. Indeed, in the literature, methods for these three problems were proposed separately given that statistical modeling for them are very different. The problem is more challenging in a longitudinal study when data are missing nonignorably. In this paper, we consider estimation in generalized linear models under these three incomplete data models. We propose a general approach based on expected estimating equations to solving these three incomplete data problems in a unified fashion. Bias analysis for naive estimation is conducted for some specific models. This expected estimating equation approach can be easily implemented and its asymptotic covariance can be obtained by sandwich estimation. Intensive simulation studies are performed under various incomplete data settings. The proposed method is applied to a longitudinal study of oral bone density in relation to body bone density.
Boosting with Missing Predictors

Abstracts

Boosting is an important tool in classification methodology. It combines the performance of many weak classifiers to produce a powerful committee, and the validity of it can be explained by additive modeling and maximum likelihood. The method has very general applications, especially to high dimensional predictors. For example, it can be applied to distinguish cancer samples from healthy control samples by using antibody microarray data. Microarray data are often high dimensional, and many of them are incomplete. One natural idea is to impute a missing variable based on observed predictors. The problem itself becomes more challenging when the missing data pattern is not monotone. In this paper, we propose an imputation method based on an iterative algorithm. This method can be applied to the situation when a complete-case subset does not even exist. Finite sample performance was examined via intensive simulation studies. We apply the method to a pancreatic cancer study in which serum-protein microarrays are used for classification.

Likelihood robustification in parametric regression with missing data

Abstracts

In parametric regression with missing data, when either the covariates are missing with or without auxiliary information, or the outcome are missing with auxiliary information, additional semiparametric or parametric models are often required for carrying out the likelihood inferences. To protect the regression parameter estimator from misspecification of the covariate models or the misspecification of the missing data mechanism model in the alternative inverse missing probability weighted approach, Robins et al. proposed a doubly robust procedure for making inferences on the regression parameter. Their approach to finding the doubly robust estimating equations was based on the projection of the weighted estimating score. A general doubly robust estimator may have low efficiency and computing the best doubly robust estimator, the locally efficient estimator, can be very challenging even for missing data with the simplest missing pattern when Robins et al.’s approach is followed. We propose an alternative representation of the semiparametric efficient score. Computationally, likelihood robustification, an approximation to the locally semiparametric efficient estimator based on the proposed representation is relatively easy to obtain and the estimator obtained from the approximating score has the doubly robust property when data are missing at random. Asymptotic inference on the regression parameter based on the likelihood robustification estimator is also proposed. Simulation results show that estimates based on the likelihood robustification performs well with finite sample sizes.
Workshop on Missing Data

主講人簡歷：

王清雲(Ching-Yun Wang) 是一名 Members of Fred Hutchinson Cancer Research Center at Seattle. 他於1993年在Texas A&M University獲得Ph.D. in Statistics學位。Ching Yun的主要專長是在測量錯誤模型、缺失數據回歸、營養流行病學和半參量回歸模型的發展和使用。他也對開發統計方法進行長期和失敗時間數據分析、偏斜樣本和相關數據感興趣。

陳華雲(HUA YUN CHEN) 是伊利諾大學芝加哥分校公共卫生學院的副教授。他在1998年於密歇根大學獲得Biostatistics學位。Hua Yun的主要專長是在缺失數據的分析、進階統計方法的發展和使用。他也對開發統計方法對於偏斜樣本和相關數據感興趣。

Welcome

聯絡人：程毅豪   yhchen@stat.sinica.edu.tw

TEL: (02) 27835611-303