Regression analysis for probabilistically linked longitudinal data

Abstract: Probabilistic data linkage is an attractive data collection option when direct measurement is impossible or extremely costly. Furthermore, the cost of linking is typically small compared to the benefit that can be derived from analysis of the linked data. One important linkage application is where different data sets relating to the same individuals at different points in time are linked to provide a longitudinal data record for each individual, thus permitting longitudinal analysis for these individuals. However, without access to the same unique identifier in each of the linked data sets, there is always the possibility that linkage errors in the merged data could lead to a longitudinal record ostensibly relating to a single individual being actually made up of a composite of data items from different individuals. This in turn could lead to bias and loss of efficiency for the longitudinal modelling process.

In this talk I will present recent results for unbiased regression inference using longitudinally linked data in the presence of probabilistic linkage errors. The approach is based on that described in Chambers (2009) and I will focus on the situation where the linked data are obtained by independently linking three data sources. Two scenarios will be investigated in detail. In the first, all three data sources represent different registers for the same population. In the second, I will extend this theory to where these data sources are a mix of register data and sample survey data.