

REGRESSION ANALYSIS WITH LINKED DATA

P. Lahiri and Michael D. Larsen
University of Maryland and Iowa State University

June 24, 2004

ABSTRACT

Record linkage, or exact matching, can be used to join together two files that contain information on the same individuals, but lack unique personal identification codes. The possibility of errors in linkage causes problems for estimating the relationships between variables on the two files. The effect is analogous to the impact of measurement error. A model of a linear regression relationship between variables in linked files is proposed. Assuming the probabilities that pairs of records are links are known, an unbiased estimator of the regression coefficients is derived. Methods for estimating the linkage probabilities by using mixture models are discussed. A consistent estimator of the covariance matrix of the proposed estimator is proposed. A bootstrap estimator is used to reflect the impact of the uncertainty in record linkage model parameters on the estimators of the regression parameters. A simulation study compares the performance of the proposed estimator and alternatives.