**SURV 622/SURVMETH 622 Fundamentals of Data Collection**

**Assignment #1: Data linkage exercise**

Due date: Monday, February 24, 2025

Using the RecordLinkage software in R, you will 1) test different specifications for determining the links between an "administrative" data set and a "survey" data set that have been created for the purpose of this exercise and 2) write a report that describes the specifications that were tested and discusses the results.

The two data files are named admindata.csv and surveydata.csv; they are available to be downloaded from the course website. In addition to a set of potential linking variables, these files also include a common record identifier that is not intended for use in linking but that can be used to determine whether a proposed link is or is not a true match. A sample program that illustrates relevant syntax for the RecordLinkage package in R will be provided and discussed in class.

You should experiment with linking specifications that vary along the following dimensions:

- Choice of linking variables
- Choice of blocking variables, if any
- Whether strings must match exactly or to some lesser tolerance as specified by a string comparison (e.g., Jaro Winkler)

You should consider 1) linking specifications with all pairs designated as either links or non-links and 2) linking specifications with an intermediate category of pairs requiring clerical review.

Some guidelines for this project:

- Be systematic in your approach. Start with a "deterministic" approach using a core set of linking variables (e.g., name and year of birth), no blocking and no string comparison (i.e., requiring that linking variables match exactly). Then explore what happens when you vary these choices—adding linking variables, adding blocking variables, using a string comparison, and, when using a string comparison, setting a higher or lower agreement threshold.
- Variables used for blocking should not have missing values. City, zip code or first letter of last name, for example, are good candidates for blocking; race or marital status are missing for a significant number of survey records and are not good choices.
- Consider *precision* and *sensitivity* as key metrics for assessing how well a given specification performed.
- In specifications that allow for clerical review, consider the number of cases assigned for clerical review.