Received: 25 June 2019 | Accepted: 28 September 2020

DOI: 10.1111/rssa.12630

ORIGINAL ARTICLE

Linkage-data linear regression

Li-Chun Zhang[1,2] | **Tiziana Tuoto**[3]

[1]University of Southampton, Southampton, UK

[2]Statistics Norway & University of Oslo, Oslo, Norway

[3]Istat, Rome, Italy

Correspondence
Li-Chun Zhang, University of Southampton, Southampton, UK; Statistics Norway & University of Oslo, Oslo, Norway.
Email: L.Zhang@soton.ac.uk

[Corrections added on 14 December 2020, after first online publication: Minor typographical errors were corrected throughout the article.]

Abstract

Data linkage is increasingly being used to combine data from different sources with the aim of identifying and bringing together records from separate files, which correspond to the same entities. Usually, data linkage is not a trivial procedure and linkage errors, false and missed links, are unavoidable. In these cases, standard statistical techniques may produce misleading inference. In this paper, we propose a method for secondary linear regression analysis, where the linked data have to be prepared by someone else, and neither the match-key variables nor the unlinked records are available to the analyst. We develop also a diagnostic test for the assumption of non-informative linkage errors, which is required for all existing secondary analysis adjustment methods. Our approach provides important advantages: it relies on the realistic assumption that the probabilities of correct linkage vary across the records but it does not assume that one is able to estimate the probability of correct linkage for each individual record. Moreover, it accommodates in a simple manner the general situation where the files are of different sizes and none of them is a subset of another. The proposed methodology of adjustment and testing is studied by simulation and applied to real data.

KEYWORDS

data integration, diagnostic test, linkage error, method of least squares, record linkage

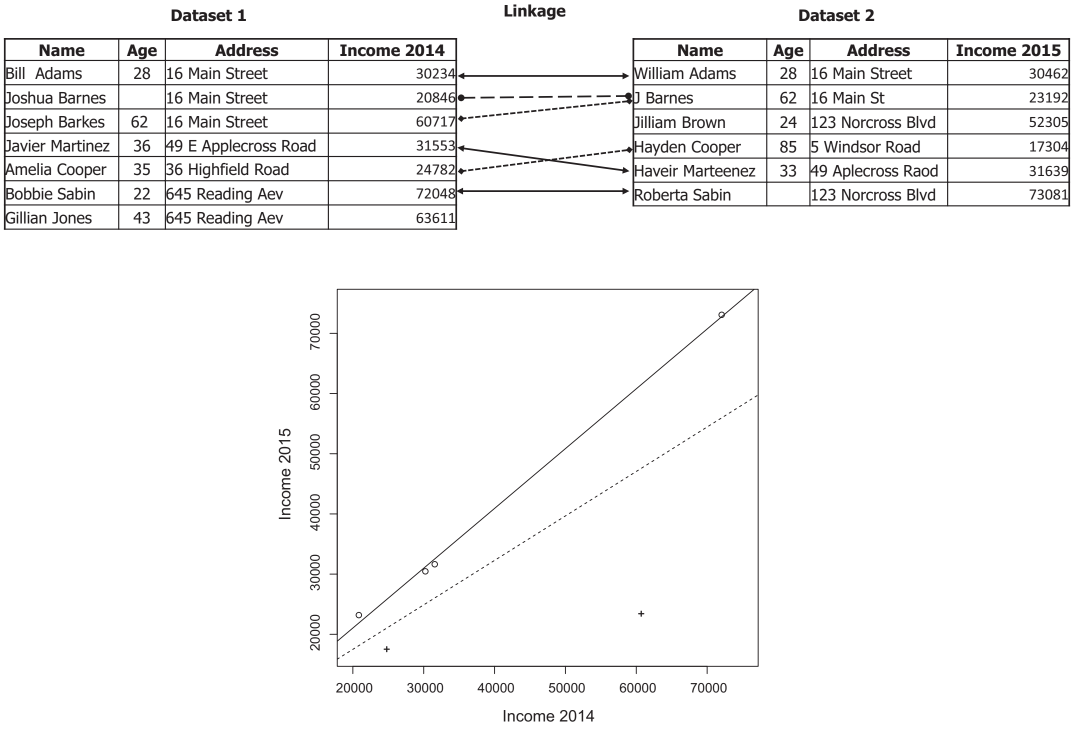
© 2020 Royal Statistical Society

1 | INTRODUCTION

Computerised record linkage is increasingly common for scientific investigation, policy analysis and commercial development, where one aims to identify and bring together the records (with associated observations) in separate data files, which correspond to the same entities or individuals (Christen, 2012; Fellegi & Sunter, 1969; Harron et al., 2015; Herzog et al., 2007). Industrial-strength applications to large population size data sets have become relatively straightforward, for example, when population census data files are linked over time to create longitudinal population data sets (Zhang & Campbell, 2012), or population-wide administrative registers are linked to create pseudo population spine in the absence of a Central Population Register (Owen et al., 2015). In epidemiology and medical studies, record linkage is extensively used in many countries to enhance data on clinical performance and patient health outcomes (e.g. Harron et al., 2016). Record linkage is a necessary step for estimating the size of hidden or hard-to-count populations, that is, illegal drug users, drinking drivers, illegal migrants, civil war victims, just to cite few examples of studies on human population (van der Heijden et al., 2014; Rosman, 2001; Seybolt et al., 2013); studies on wild animal populations provide plenty of application (Creel et al., 2003; Link et al., 2010; McClintock et al., 2014; Wright et al., 2009). In our illustrative application in Section 4, we consider linked income data from tax registers in two consecutive years, and linear regression of year-on-year incomes for a simple analysis of the development at local (municipality) level. Using administrative data here allows for disaggregated analysis that otherwise cannot be supported by survey sampling, because of the limited sample size and the fact that income may be considered 'sensitive', which causes non-response and/or under-reporting errors in surveys.

When there does not exist a unique identifier that allows for exact matching, record linkage is performed using soft identifiers, the so-called *key variables*, such as name, age, address, etc. Let each pairing of records of the same entity be a *match*. Let each pairing of records that results from record linkage be a *link*. Insofar as the key variables may be affected by measurement errors, *linkage errors* are unavoidable, so that the links may not be identical to the matches. There are two types of linkage error: either the linked records do not actually refer to the same entity, or if one fails to link the records that refer to the same entity. Figure 1 provides an illustration using fictive individuals and income data, where there are three correct *links* (solid), two errors of *false linkage* (dashed) and one of *missing match* (long-dashed). The plot shows the ordinary least squares fit (solid line) based on the four unknown matches (circle) and that (dashed) based on the five observed links ('+' for the two incorrect links). Clearly, treating the linked data set as if it were true generally causes bias of the resulting analysis. For a situation like the one in Figure 1, one needs to deal with *at least* three problems.

1. Different individuals (or entities) can have different probabilities for being incorrectly linked or missed (given a match exists), which we refer to as the problem of *heterogeneous linkage errors*.
2. There are *unmatched* individuals in *both* files that cannot possibly be correctly linked, which we refer to as the problem of *incomplete match space*. In Figure 1 these are Barkes, A. Cooper and Jones in file 1, and Brown and H. Cooper in file 2. *Complete match space* would have been the case here had none of these unmatched individuals existed, or if they had only existed in *one* of the two files, say, when file 1 is a sample taken from file 2.
3. Whether (Joseph Barnes, J. Barnes) are a match is a mutually exclusive event of whether (Joseph Barkes, J. Barnes) are a match, as long as there are no duplicated records in each file, which we refer to as the problem of *linkage data structure*. Due to the linkage data structure, it would, for example, be wrong to model (Joseph Barnes, J. Barnes)'s being a match as a Bernoulli event that is statistically independent of (Joseph Barkes, J. Barnes)'s being a match.



Dataset 1

Name	Age	Address	Income 2014
Bill Adams	28	16 Main Street	30234
Joshua Barnes		16 Main Street	20846
Joseph Barkes	62	16 Main Street	60717
Javier Martinez	36	49 E Applecross Road	31553
Amelia Cooper	35	36 Highfield Road	24782
Bobbie Sabin	22	645 Reading Aev	72048
Gillian Jones	43	645 Reading Aev	63611

Dataset 2

Name	Age	Address	Income 2015
William Adams	28	16 Main Street	30462
J Barnes	62	16 Main St	23192
Jilliam Brown	24	123 Norcross Blvd	52305
Hayden Cooper	85	5 Windsor Road	17304
Haveir Marteenez	33	49 Aplecross Raod	31639
Roberta Sabin		123 Norcross Blvd	73081

FIGURE 1 Fictive income data 2014–2015. Top: links between the records, including correct link (——), false link (- - -) and missing match (- -). Bottom: linear regression, solid line based on unknown matches (o), and dashed line based on observed links (+)

1.1 | Related works

The awareness of misleading inference from standard statistical techniques in the presence of linkage errors dates back to Neter et al. (1965). Linear regression is studied by Scheuren and Winkler (1993, 1997) and Lahiri and Larsen (2005), where the data analyst and the linker are essentially the same. Chambers (2009) and Chambers and Diniz da Silva (2020) adopt the perspective of secondary analysts, who have no access to the key variables and the separate data files, nor the detailed knowledge or tools to replicate the actual linkage procedure (Zhang, 2019). Consequently, Chambers (2009) adopts a greatly simplifying assumption, referred to as the *exchangeable linkage error (ELE)* model, where there exists a constant false linkage probability and mismatching is completely random in the case of false linkage. While the ELE assumption is practically appealing, it cannot properly accommodate heterogeneous linkage errors. Moreover, as we shall explain in more details in Section 2, the ELE model is only applicable if one treats any incomplete match space as if it were complete. But the false linkage error of an unmatched individual (such as Brown in Figure 1) always has probability one, so it cannot be the same as that of a matched individual (such as Martinez) who can be linked correctly.

Nearly all the frequentist methods for the analysis of linked data are based on the *linkage model* of the probability that a record in one data set is linked to *each* of the records in the other. The ELE model is the simplest linkage model. Techniques such as regression analysis, estimation equation and analysis of contingency tables are studied by Scheuren and Winkler (1993, 1997), Lahiri and Larsen (2005), Chambers (2009), Chipperfield et al. (2011), Hof and Zwinderman (2012), Kim and

Chambers (2012a, b), Chipperfield and Chambers (2015), Han and Lahiri (2018) and Enamorado et al. (2019). Again, as we shall explain in Section 2, in reality the linkage model cannot cope with incomplete match space, even when the ELE assumption is relaxed to accommodate heterogeneous linkage errors. Yet incomplete match space is generally the case when data originate from different sources, such as when linking hospital patient records to welfare payment records. It is fundamentally different to the situation, where one set of individuals form a sample of the other set (i.e. population), where there are no individuals in the sample who cannot possibly be linked correctly.

Bayesian inference is based on the posterior distribution of the unknown set of matched entities. Different modelling approaches are used for the linkage key variables that are subjected to measurement errors, for example, Tancredi and Liseo (2011) and Stoerts et al. (2017) extend the hit-miss model of Copas and Hilton (1990), whereas Sadinle (2014, 2017) models the comparison vector of key variables following the Fellegi and Sunter (1969) tradition. See also Gutman et al. (2013) for a modelling approach, which includes both variables subjected to measurement errors and others that do not. However, it is common that the variables being modelled for linkage are inaccessible to the secondary analyst. Handing out multiple posterior sets of matched entities may be impractical, together with the associated variables needed for analysis, especially if the analysis requires a large number of posterior draws. Although there are improvements in the direction of scalability (Marchant et al., 2019), there still does not exist any reported Bayesian linkage application to files of the size of a population census.

Goldstein et al. (2012) and Gutman et al. (2015) apply multiple imputation techniques to analysis of linkage data, which do not handle the problem of linkage data structure like the other Bayesian methods above. Restrictions due to linkage data structure are not built into these imputation methods.

1.2 | Outline of the paper

In this paper, we consider *linkage-data linear regression*, where one aims to estimate the regression coefficients *only* based on the linked data set. In particular, we adopt the *secondary* analyst perspective, where the linked data have to be prepared by someone else; neither the unlinked records nor the key variables in the separate files are available to the analyst. We develop a novel frequentist method of *Pseudo Ordinary Least Squares (OLS)*, which deals with all the three problems exemplified above in Figure 1, that is, heterogeneous linkage errors, incomplete match space and linkage data structure. Like all the methods referenced in this Introduction, the key assumption to our approach is that the linkage errors are non-informative of the regression model parameters. The assumption will be defined and discussed in Section 2. Moreover, for the first time we shall construct an accompanying diagnostic test for the non-informative linkage error assumption, which can provide helpful guidance in practice. Application to real income data and simulation studies suggest that the assumption can be met at least approximately in many situations, and the Pseudo-OLS estimator is more efficient than the existing methods in the cases of incomplete match space that are examined here.

The rest of the paper is organised as follows. In Section 2 we start by introducing the basic notations and the set-up of linkage-data linear regression. In Section 2.1, we recall the existing frequentist methods and explain carefully why they do not fully meet the challenges of incomplete match space. Section 2.2 defines and discusses the non-informative linkage error assumption. Our proposed approach is then developed in Sections 2.3, 2.4 and 2.5, including the underlying assumptions and the consistency of the resulting regression coefficient estimator. Section 2.6 analyses the bias of the existing methods, which arises from treating the incomplete match space as if it were complete. In Section 3 we develop a diagnostic test for the non-informative linkage error assumption. An application to linked income data from tax registers is given in Section 4, which demonstrates considerable efficiency gains by our method against the existing ones. We carry out a simulation study in Section 5, which helps us to better appreciate the application results and to explore some other aspects of the proposed methodology of adjustment and testing. We conclude with some brief remarks in Section 6.

2 | METHODS

Let $y_i = x_i^\top \beta + \epsilon_i$ be a linear regression model, where x_i is the $p \times 1$ vector of covariates, and β is the parameter of interest. Let data set D_1 contain the covariates x_i for record $i \in D_1$, and let data set D_2 contain the dependent variable y_j for $j \in D_2$. We assume that duplicated records have been successfully removed from both. Let $N_1 = |D_1|$ and $N_2 = |D_2|$ be the sizes of D_1 and D_2. Let D_M be the set of matched entities between D_1 and D_2, that is, those ones that can possibly be correctly linked, to which the linear regression model applies. Let

$$\Omega = D_1 \times D_2 = M \cup U,$$

where $M = \{ (i, i) : i \in D_M \}$ contains the matches, and U contains all the mismatched pairs of records. Let $N_M = |M|$ be the size of M. In the ideal case, one would estimate β based on the pairs of records in M. However, M is unknown. Suppose a record linkage procedure yields the set of links, between records in D_1^* from D_1 and D_2^* from D_2, respectively, denoted by

$$M^* = \{ (i,j) : i \in D_1^*, j \in D_2^* \},$$

where $N^* = |D_1^*| = |D_2^*| = |M^*| \leq min(N_1, N_2)$, and $M^* \neq M$ whenever linkage errors are present. In linkage-data linear regression one aims to estimate β only based on the linked data set, which can take on any of the following expressions in this paper:

$$(x, y)_{M^*} = \{ (x_i, y_j) : (i,j) \in M^* \} = \{ (x_i, y_i^*) : y_i^* = y_j, (i,j) \in M^* \}.$$

Let $D_{1M}^* = D_1^* \cap D_M$ be the set of matched entities in D_1 that are linked, and $D_{2M}^* = D_2^* \cap D_M$ those from D_2. Let D_{MM}^* be the set of correctly linked entities, where $\{ (i, i) : i \in D_{MM}^* \} = M^* \cap M$. Let $N_{MM}^* = |D_{MM}^*|$ be its size. We have $D_{MM}^* \subseteq D_{1M}^* \subseteq D_1^*$ and $D_{MM}^* \subseteq D_{2M}^* \subseteq D_2^*$.

For an illustration using Figure 1, let file 1 contain $D_1 = \{ 1, 2, 3, 4, 5, 6, 7 \}$ and let file 2 contain $D_2 = \{ 1, 2, 8, 9, 4, 6 \}$, both in the running order from top to bottom, where $D_M = \{ 1, 2, 4, 6 \}$ are the matched individuals and {3,5,7,8,9} are the unmatched ones. We have $D_1^* = \{ 1, 3, 4, 5, 6 \}$ and $D_{1M}^* = \{ 1, 4, 6 \}$, $D_2^* = \{ 1, 2, 9, 4, 6 \}$ and $D_{2M}^* = \{ 1, 2, 4, 6 \}$. The set of links is $M^* = \{ (1, 1), (3, 2), (4, 4), (5, 9), (6, 6) \}$. The correctly linked individuals can only come from D_M, which are $D_{MM}^* = \{ 1, 4, 6 \}$.

2.1 | Two linkage-model estimators for complete match space

Consider the case of complete match space, where $N \equiv N_1 = N_2 = N_M$. Suppose each record in D_1 is linked to one and only one record in D_2, such that $(N^*, D_1^*, D_2^*) = (N, D_1, D_2)$. The linked y-value for $i \in D_1^*$ is $y_i^* = \sum_{j \in D_2^*} a_{ij} y_j$, where $a_{ij} = 1$ if $i \in D_1^*$ is linked to $j \in D_2^*$ and $a_{ij} = 0$ otherwise. Notice that i and j refer to distinctive records themselves, regardless how they appear or are arranged in the two files. False linkage of $i \in D_1^*$ is the case if $a_{ij} = 1$ for $j \in D_2^*$ and $j \neq i$. However, a_{ij} is unobserved, since the true matches are unknown. What is observed is whether or not $i \in D_1^*$, that is, record $i \in D_1$ is linked or not, denoted by $\ell_i = 1$ or $\ell_i = 0$. In the special setting here, we have $\ell_i = 1$ for all $i \in D_1$. Denote the conditional expectation of a_{ij} given linkage,by

$$p_{ij} = E(a_{ij} \,|\, \ell_i = 1) = Pr(a_{ij} = 1 \,|\, \ell_i = 1).$$

Let $P_{N\times N} = [p_{ij}]$ be the matrix of p_{ij}'s. Let $X_{N\times p}$ be the covariate matrix associated with D_1, and $y_{N\times 1}$ the dependent vector of D_2, in the matched ordering such that the diagonal of P corresponds to M. Let $y^*_{N\times 1}$ be the vector of linked y-values, which is a linear transformation of y via $[a_{ij}]$.

Provided the linkage indicators $[a_{ij}]$ are independent of $(x, y)_M$, we have

$$E(y^* \,|X, y) = Py.$$

Given complete match space, the regression model applies to all the units in D_2, so that $E(y|X) = X\beta$. Thus, $E(y^* \,|X) = Z\beta$ for $Z = PX$. Lahiri and Larsen (2005) propose OLS fit:

$$\hat{\beta}_{LL} = (Z^{\mathsf{T}}Z)^{-1}Z^{\mathsf{T}}y^*.$$

Chambers (2009) notices in addition an unbiased *adjusted least squares* fit:

$$\hat{\beta}_A = (X^{\mathsf{T}}PX)^{-1}X^{\mathsf{T}}y^*$$

The matrix P does not contain sensitive information and, in theory, could be supplied by the data linker. In practice, however, there is currently a lack of consensus on how to estimate the matrix P. See discussions of alternative approaches in Lahiri and Larsen (2005), Han and Lahiri (2018), Chipperfield and Chambers (2015), and Tuoto (2016). Moreover, these methods require access to the key variables, which is only possible for the data linker. Chambers (2009) proposes the ELE model of P, where

$$p_{ii} = \lambda \text{ and } p_{ij} = (1 - \lambda)\,/\,(N - 1). \tag{1}$$

which ignores the problems of heterogeneous linkage errors. Even when the model (1) is relaxed to accommodate heterogenous linkage errors with varying p_{ij}'s, the linkage-model approach still cannot cope with the problem of incomplete match space. In this paper, we propose linear regression methods which makes it unnecessary for the data linker to supply the matrix P.

Again, take the example in Figure 1 and consider Adams ($i = 1$) and Barkes ($i = 3$). The expectation of their linked y-value, respectively, are given as

$$E(y_1^* \,|X, y, \ell_1 = 1) = p_{11}y_1 + p_{12}y_2 + p_{18}y_8 + p_{19}y_9 + p_{14}y_4 + p_{16}y_6,$$
$$E(y_3^* \,|X, y, \ell_3 = 1) = p_{31}y_1 + p_{32}y_2 + p_{38}y_8 + p_{39}y_9 + p_{34}y_4 + p_{36}y_6,$$

provided non-informative linkage errors. Since Adams is a matched individual that can be linked correctly, one can, for example, let $p_{11} = \lambda_1$ and $p_{1j} = (1 - \lambda_1)\,/\,4$, for any other $j \in D_2^*$, given that the secondary analyst only sees the five links that are provided. But this would mean to assume that the unlinked individual Brown ($i = 8$) in $D_2 \setminus D_2^*$ has no chance of being linked with Adams, that is, treating the incomplete match space as if it were complete. Next, since Barkes is an unmatched individual, it would be totally wrong to act similarly, because there is no record at all in D_2 for Barkes. One might consider setting $p_{3j} \equiv 1\,/\,5$ as an assumption of random false linkage. However, without knowing the true matched or unmatched status of Adams and Barkes, one would not know if p_{1j}'s or p_{3j}'s should be assigned. This shows that the linkage-model approach cannot cope with incomplete match space.

Thus, in reality, one can only apply the ELE model (1), by assuming that the linked sets (D_1^*, D_2^*) form complete match space in any case. Clearly, this is not satisfactory conceptually: although one may assume $y_3 = x_3^\top \beta + \epsilon_3$ for Barkes in D_1^*, one would not find y_3 among $y^* = (y_1, y_2, y_9, y_4, y_6)^\top$. Similarly, although one may assume that there exists x_9 for Cooper in D_2^*, such that $y_9 = x_9^\top \beta + \epsilon_9$, one would not find x_9 in $X_{D_1^*}$. However, as we will discuss later in Section 2.6, doing so may still yield useful bias reduction compared to the *face-value* OLS, given by

$$\hat{\beta}^* = (X_{D_1^*}^\top X_{D_1^*})^{-1} X_{D_1^*}^\top y^*.$$

For now we only notice some intuition why this may be the case. Provided the false linkage rate is low, $(1 - \lambda)/N^* \approx (1 - \lambda)/N_2 \approx 0$ for large N^*, and the misspecification of p_{ij}, where $i \neq j$, may not matter much for the records D_M. Moreover, the proportion of unmatched but linked entities is then also low, so that there are relatively few rows like that for Barkes here. In short, the effects due to the misspecification of the P-matrix may be limited given *low* false linkage rate, and the linkage-model estimators $\hat{\beta}_{LL}$ and $\hat{\beta}_A$ may still remove most of the bias of the face-value estimator $\hat{\beta}^*$.

2.2 | Non-informative linkage error assumption

The linkage model essentially requires one to specify, for any given record i in D_1, the probability of $a_{ij} = 1$ for *all* the records $j \in D_2$. To accommodate incomplete match space and heterogeneous linkage errors, we specify the *non-informative linkage error (NILE)* assumption as follows:

$$\lambda_i = \Pr(a_{ii} = 1 \mid \ell_i = 1, X, y) = \begin{cases} \Pr(a_{ii} = 1 \mid \ell_i = 1) & \text{for } i \in D_M \\ 0 & \text{for } i \notin D_M \end{cases} \tag{2}$$

and, for $i \in D_1$ (or D_2), the probability of linkage is independent of (X,y), that is,

$$\psi_i = \Pr(\ell_i = 1 \mid X, y) = \Pr(\ell_i = 1). \tag{3}$$

Heterogeneous linkage error is the case if λ_i varies over D_M and ψ_i over D_1 (or D_2). The assumption (2) accommodates incomplete match space, assigning zero chance of correct link to any unmatched entities in $D_1 \setminus D_M$ or $D_2 \setminus D_M$, without needing to specify p_{ij} for $i \in D_1$, $j \in D_2$ and $j \neq i$. It is possible to incorporate in ψ_i a sample inclusion probability, as when D_1 is a sample from population D_2.

We introduce also a slightly weaker NILE assumption as follows, which we use for the consistency results later on. Let z_i be a well-defined function of x_i and y_i, such as $x_i y_i$ for $i \in D_M$ or $z_i = x_i x_i^\top$ for $i \in D_1$, where D_z is the corresponding entity set of z_i, which is of the size N_z. Let $\overline{\psi} = \sum_{i \in D_z} \psi_i / N_z$, $\overline{z} = \sum_{i \in D_z} z_i / N_z$, and $S_{\psi z} = \sum_{i \in D_z} (\psi_i - \overline{\psi})(z_i - \overline{z}) / N_z$. *Asymptotic NILE over* D_z is the case, as $N_z = |D_z| \to \infty$, provided (2) and

$$S_{\psi z} \to 0, \tag{4}$$

that is, ψ_i and z_i are empirically uncorrelated over the set D_z. Notice that (X,y) can be treated as constants in (4), to be incorporated in a design-based approach to sample survey data, where record linkage is needed. The assumption (4) is weaker than (3), since (3) implies (4), but not *vice versa*.

Since regression analysis is conditional on X, other authors using the linkage-model approach assume non-informative linkage error is the case if a_{ij}'s are independent of y conditional on X (Chambers, 2009; Lahiri & Larsen, 2005). While the formulation is parsimonious, in reality it is not weaker than the definition here, as we discuss below. Let c_i be the linkage key variables, and $c_i^{(1)}$ the observed value of c_i in D_1 and $c_i^{(2)}$ that in D_2. In many applications, c_i is not involved in the regression, such as when c_i consists of Name, Date of Birth and Address. It seems reasonable to assume that the potential measurement errors affecting $(c_i^{(1)}, c_i^{(2)})$ are independent of (x_i, y_i) given c_i. Let C, $C^{(1)}$, $C^{(2)}$ be the collections of $c_i, c_i^{(1)}, c_i^{(2)}$, respectively, we would then have

$$\Pr(a_{ii} = 1, \ell_i = 1 \mid X, y, C, C^{(1)}, C^{(2)}) = \Pr(a_{ii} = 1, \ell_i = 1 \mid C, C^{(1)}, C^{(2)}),$$

so that (λ_i, ψ_i) neither depend on y nor X, either conditional on $(C, C^{(1)}, C^{(2)})$ or after integrating out $(C, C^{(1)}, C^{(2)})$ with respect to whichever distribution they have.

It is still possible sometimes that a key variable, which necessarily is present in *both* data sets, may be related to the x-variables, but not the y-variable. For example, Age or Country of Birth may form part of x_i, possibly after some regrouping. Let x_{ic} contain these common variables between c_i and x_i. Let x_i^R be the remaining x-variables, and c_i^R the remaining key variables. The NILE assumption is satisfied provided x_{ic} is used as blocking variables in record linkage, such that only records within the same block can possibly be linked to each other, because the blocking variables are considered to be free of measurement errors. This is typically the case with the variables Age and Country of Birth.

However, it is conceivable that the overlapping x_{ic} is not used as a blocking variable. It is currently an 'open question' (Chambers & Kim, 2015) how to deal with informative linkage errors. The problem is complicated not least when the observed values $(x_{ic}^{(1)}, x_{ic}^{(2)})$ may differ from the true x_{ic} and, depending on the method of record linkage, $x_{ic}^{(1)}$ may or may not be equal to $x_{ic}^{(2)}$ given $\ell_i = 1$. Thus, the value of x_{ic} to be used in the linkage-data linear regression may be subject to measurement error, whether or not record i is correctly linked. In this paper, we shall assume that the potential linkage error due to such key variable covariates is negligible compared to the rest key variables c_i^R, so that the NILE assumption remains acceptable. The same is needed when non-informativeness is defined conditionally given X.

2.3 | OLS based on Gold linkage

For the first estimator of β to be considered, we assume the linked set is such that missing match is possible but not false links, to be referred to as a *Gold* linkage procedure. Denote by $D_G^* = D_1^* = D_2^*$ the Gold linkage set, which involves a further selection from all the links that otherwise might have been considered acceptable. Linkage procedures that allow false links are referred to as *sub-Gold* linkage. The terms Gold and sub-Gold are only used as shorthands of the two record linkage settings, and no emotive connotation is intended. We have $\lambda_i = \Pr(a_{ii} = 1 \mid \ell_i = 1) = 1$ by Gold linkage. Denote by $\tilde{\beta}$ the ideal OLS based on $(x, y)_M$, and by $\widehat{\beta}_G$ the OLS based on $(x, y)_{D_G^*}$, which are, respectively,

$$\begin{aligned} \tilde{\beta} &= \left(\sum_{i \in D_M} x_i x_i^\top \right)^{-1} \left(\sum_{i \in D_M} x_i y_i \right), \\ \widehat{\beta}_G &= \left(\sum_{i \in D_G^*} x_i x_i^\top \right)^{-1} \left(\sum_{i \in D_G^*} x_i y_i^* \right) = \left(\sum_{i \in D_G^*} x_i x_i^\top \right)^{-1} \left(\sum_{i \in D_G^*} x_i y_i \right). \end{aligned} \tag{5}$$

Proposition 1 *Asymptotically, as* $N_M = |M| \to \infty$*, we have* $\widehat{\beta}_G - \tilde{\beta} \xrightarrow{P} 0$*, provided*

(g1) NILE assumption (2), with $\lambda_i \equiv 1$, and (4) over D_M,
(g2) $E(N^*_G / N_M) \to \psi > 0$, where $N^*_G = |D_{G^*}|$.

Under the regression model, the variance of $\widehat{\beta}_G$ conditional on $X_{D^*_G}$ is given by

$$V(\widehat{\beta}^*_G) = (X^\top_{D^*_G} X_{D^*_G})^{-1} (X^\top_{D^*_G} V(y_{D^*_G}) X_{D^*_G}) (X^\top_{D^*_G} X_{D^*_G})^{-1}.$$

The convergence can be established directly under (g1) and (g2), where the x- and y-values are treated as constants. For any $z_i = z(x_i, y_i)$ for $i \in D_M$, we have

$$E\left(\sum_{i \in D^*_G} z_i | X, y \right) = E\left(\sum_{i \in D_M} \ell_i z_i | X, y \right) = \sum_{i \in D_M} E(\ell_i | X, y) z_i = \sum_{i \in D_M} \psi_i z_i$$

Thus, ℓ_i being an unbiased estimator of ψ_i, we have $\sum_{D^*_G} z_i / N_M - \overline{\psi}\,\overline{z} \xrightarrow{P} 0$, provided (g1). Provided (g2), so that $\overline{\psi} \to \psi$, we have $\sum_{D^*_G} z_i / N^*_G - \overline{\psi}\,\overline{z} \xrightarrow{P} 0$. The result $\widehat{\beta}_G - \tilde{\beta} \xrightarrow{P} 0$ follows from replacing z_i with $x_i x_i^\top$ and $x_i y_i$ in both the estimators.

We notice that the consistency of $\widehat{\beta}_G$ given by (5) holds when record linkage follows sampling from D_1 or D_2 or both, provided sampling is non-informative of the x- and y-values in D_M. Finally, in the case of $V(y_{D_M}) = \sigma^2 I_{N_M \times N_M}$, $V(\widehat{\beta}^*_G)$ reduces to $(X^\top_{D^*_G} X_{D^*_G})^{-1} \sigma^2$. The relative efficiency to the ideal $\tilde{\beta}$ converges to $1/\psi$, as $N_M \to \infty$ asymptotically.

2.4 | Covariance of (x_i, y^*_i)

To estimate β based on sub-Gold linkage, we shall make use of the covariance between x_i and its linked y-value. The result below holds for any analysis of interest, not just linear regression. For any $i \in D_1$, we observe x_i. At most one link is allowed for each record. For any linked record $i \in D^*_1$, its linked y-value is given by $y^*_i = \sum_{j \in D_2} a_{ij} y_j$. Provided NILE (2), for any $i \in D_M$, we have

$$\begin{aligned} Cov(x_i, y^*_i \,|\, \ell_i = 1) &= Cov(x_i, a_{ii} y_i \,|\, \ell_i = 1) + \sum_{j \neq i} Cov(x_i, a_{ij} y_j \,|\, \ell_i = 1) \\ &= E(a_{ii} \,|\, \ell_i = 1) Cov(x_i, y_i) + \sum_{j \neq i} E(a_{ij} \,|\, \ell_i = 1) Cov(x_i, y_j). \end{aligned}$$

As long as x_i and y_j are uncorrelated for $i \neq j$, we have $Cov(x_i, y^*_i \,|\, \ell_i = 1, a_{ii} = 1) = Cov(x_i, y_i)$ given correct linkage, and $Cov(x_i, y^*_i \,|\, \ell_i = 1, a_{ii} = 0) = 0$ given false link of any matched entity $i \in D_M$, or linkage of an unmatched unit $i \in D_1 \setminus D_M$. It follows that, for any $i \in D_1$,

$$Cov(x_i, y^*_i \,|\, \ell_i = 1) = \lambda_i Cov(x_i, y_i),$$

where λ_i is given by Eq. (2). That is, false links on average move the observed covariance among the linked pairs of records towards zero. Moreover, to account for the effective *matched* sample size of the empirical covariance between x_i and y_i^* over the linked set D_1^*, one only needs to know the total number of correct matches in D_1^*, but not necessarily the individual λ_i's. The idea is developed below.

2.5 | Pseudo-OLS based on sub-Gold linkage

Given any sub-Gold linkage procedure, let the Pseudo-OLS fit of β be given by

$$\widehat{\beta}_P = \left(\frac{1}{N^*} X_{D_1^*}^\top X_{D_1^*}\right)^{-1} \left(\bar{x}\bar{y}^* + \widehat{\lambda}^{-1} S_{xy*}\right) \tag{6}$$

$$= \widehat{\lambda}^{-1}\widehat{\beta}^* - \left(\widehat{\lambda}^{-1} - 1\right)\left(\frac{1}{N^*} X_{D_1^*}^\top X_{D_1^*}\right)^{-1} \bar{x}\bar{y}^*, \tag{7}$$

where $\bar{x} = \sum_{i \in D_1^*} x_i / N^*$, and $\bar{y}^* = \sum_{i \in D_1^*} y_i^* / N^*$, and $S_{xy^*} = \sum_{i \in D_1^*} (x_i - \bar{x})(y_i^* - \bar{y}^*)/N^*$, and $\widehat{\lambda}$ is an estimate of the proportion of correct matches among the actual links. Notice that $\widehat{\lambda}$ can be obtained for the realised D_1^*, for example, by auditing a sample of the links in it. The expression (6) reveals that the Pseudo-OLS is based on a linkage error adjustment of the observed covariance between x_i and y_i^* in the linked data set, whilst the expression (7) shows it as a linear adjustment of the naïve face-value OLS $\widehat{\beta}^* = (X_{D_1^*}^\top X_{D_1^*})^{-1} X_{D_1^*}^\top y^*$.

Example 1 For simple linear regression $y_i = \alpha + \beta x_i + \epsilon_i$, the Pseudo-OLS is given by

$$\begin{bmatrix} \widehat{\alpha}_P \\ \widehat{\beta}_P \end{bmatrix} = \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \frac{1}{N^*}\sum_{i \in D_1^*} x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \bar{y}^* \\ \bar{x}\bar{y}^* + \widehat{\lambda}^{-1} S_{xy^*} \end{bmatrix}$$

$$\Rightarrow \widehat{\beta}_P = \widehat{\lambda}^{-1} \frac{S_{xy^*}}{S_x^2} = \widehat{\lambda}^{-1} \frac{\sum_{i \in D_1^*} (x_i - \bar{x})(y_i^* - \bar{y}^*)}{\sum_{i \in D_1^*} (x_i - \bar{x})^2} \quad \text{and} \quad \widehat{\alpha}_P = \bar{y}^* - \bar{x}\widehat{\beta}_P,$$

where $\widehat{\beta}_P$ is a multiplicative adjustment of the face-value OLS of the slope *away from 0*, for $\widehat{\lambda} < 1$. This is intuitive because, given a false link is made for $i \in D_1^*$, the face-value covariance $(x_i - \bar{x})(y_i^* - \bar{y}^*) = (x_i - \bar{x})(y_j - \bar{y}^*)$ has approximately expectation zero, as long as x_i and y_j are uncorrelated for $j \neq i$. So the face-value estimate of the slope is biased towards 0. To adjust for the bias, notice that the effective sample size underlying the linked sample covariance S_{xy^*} is just the number of true matches among the links, which is estimated by $\widehat{\lambda}N^*$. This is the basic idea underlying the Pseudo-OLS (6).

2.5.1 | Consistency conditions for Pseudo-OLS

Given sub-Gold linkage, we have $E(N_{MM}^* | D_{1M}^*) = \sum_{i \in D_{1M}^*} \lambda_i = \sum_{i \in D_1^*} \lambda_i = E(N_{MM}^* | D_1^*)$, because $\lambda_i = 0$ for the unmatched entities in $D_1^* \setminus D_{1M}^*$. We have $\widehat{\beta}_P - \tilde{\beta} \to 0$, if the difference between each term in (6) and its counterpart in $\tilde{\beta}$ converges to zero in probability. In addition to the NILE assumption and the consistency of $\widehat{\lambda}$, regularity conditions (p0.1)–(p0.3) are needed regarding the values of (x,y) associated with the unknown entities underlying D_1 and D_2. Condition (p0.1) states that x and y associated with different entities are uncorrelated with each other, such as common for linear regression given perfectly matched data set. Conditions (p0.2) and (p0.3) mean that the unmatched entities do not have pathological x or y-values, such as outliers of arbitrary magnitude. All the conditions are given below in Proposition 2, the proof of which is given in Appendix A.

Proposition 2 *Asymptotically, as* $N_M = |M| \to \infty$, *we have* $\widehat{\beta}_P - \tilde{\beta} \xrightarrow{P} 0$, *provided*

(p0.1) $Cov(x_i, y_j) = 0$ for $j \neq i$, $i \in D_1$ and $j \in D_2$,
(p0.2) $\sum_{i \in D_M} x_i / N_M - \sum_{i \in D_1} x_i / N_1 \to 0$,
(p0.3) $\sum_{j \in D_M} y_j / N_M - \sum_{j \in D_2} y_j / N_2 \to 0$,
(p1) NILE assumption (2) and (4), where (4) holds over D_1 as well as D_2,
(p2) $E(N^*) \to \infty$, and $E(N^*_{MM}/N^*) \to \lambda > 0$, and $\widehat{\lambda} \xrightarrow{P} \lambda$.

2.5.2 | Variance estimation

It is impractical to allow heterogeneous variance of ϵ_i, because we do not know the x-values in the case of a false link. We shall, therefore, assume $V(y_i) = \sigma^2$ for all $i \in D_2$. Provided NILE, it is natural to condition on the realised N^*. Given false link of $i \in D^*_1$, we have $y^*_i = y_j$, for some $j \in D_2$ and $j \neq i$, where the record j may or may not belong to D_M. In the case of $j \notin D_M$, we shall assume that there nevertheless exists a vector x_j under the regression model, even though $j \notin D_1$. Thus, we shall condition on $(X_{D_1}, X_{D_2 \setminus D_M}, N^*)$ throughout the following. We have

$$V(\widehat{\beta}_P) = \left(\frac{1}{N^*} X^\top_{D^*_1} X_{D^*_1}\right)^{-1} V(\overline{x y}^* + \widehat{\lambda}^{-1} S_{xy^*}) \left(\frac{1}{N^*} X^\top_{D^*_1} X_{D^*_1}\right)^{-1}.$$

Now, given the linkage matrix $A = [a_{ij}]$, where at most one link is allowed for a record, $y^*_i = y_j$ is conditionally independent of $y^*_k = y_l$ for $i \neq k$, since $j \neq l$ regardless if (i,j) and (k,l) are true matches or not. Thus, we have

$$V(\overline{x y}^* + \widehat{\lambda}^{-1} S_{xy^*}) = V(\overline{x y}^*) + V(\widehat{\lambda}^{-1} S_{xy^*}),$$

since $Cov(\bar{y}^*, y^*_i - \bar{y}^* | A) = 0$, hence $Cov(\bar{y}^*, \widehat{\lambda}^{-1} S_{xy^*} | A) = 0$ and $Cov(\bar{y}^*, \widehat{\lambda}^{-1} S_{xy^*}) = 0$. By working out $V(\overline{x y}^*)$ and $V(\widehat{\lambda}^{-1} S_{xy^*})$—see Appendix B for details, we obtain

$$V(\widehat{\beta}_P) \approx (X^\top_{D^*_1} X_{D^*_1})^{-1} \sigma^2 + \left(\frac{1}{N^*} X^\top_{D^*_1} X_{D^*_1}\right)^{-1} \Delta \left(\frac{1}{N^*} X^\top_{D^*_1} X_{D^*_1}\right)^{-1}, \tag{8}$$

where

$$S_{xx} = \frac{1}{N^*} \sum_{i \in D^*_1} (x_i - \bar{x})(x_i - \bar{x})^\top \quad \text{and} \quad \Delta = \left(\frac{1}{\lambda^2} - 1\right) \frac{\sigma^2}{N^*} S_{xx} + V(\widehat{\lambda}) S_{xx} \beta \beta^\top S^\top_{xx}.$$

Clearly, linkage errors cause a loss of efficiency, since the first term on the right-hand side of (8) would have been the variance had all the links been true matches and adjustment not needed. The extra variance depends on Δ, which has two contributing terms: one due to the smaller effective sample size N^*_{MM} compared to the face-value sample size N^*, the other due to the estimation uncertainty of the adjustment factor $\widehat{\lambda}$. Compared to $\widehat{\beta}_G$ by Gold linkage, the first term of (8) is smaller than $V(\widehat{\beta}_G)$, since $D^*_G \subset D^*_1$. However, the extra uncertainty in (8) due to Δ may still possibly cause loss of efficiency of sub-Gold linkage compared to Gold linkage. The matter is explored empirically in Section 5.

For plug-in variance estimation, we need an estimate of σ^2, in addition to $\widehat{\beta}_P$ and $\widehat{\lambda}$. Applying the standard formula of OLS variance estimator to the linkage data, we obtain

$$S_{ee}^* = \frac{1}{N^* - p} \sum_{i \in D_1^*} (y_i^* - \widehat{\beta}_P^\top x_i)^2 = \frac{1}{N^* - p} \sum_{i \in D_1^*} [\,(y_{j_i} - \widehat{\beta}_P^\top x_{j_i}) - \widehat{\beta}_P^\top (x_i - x_{j_i})\,]^2,$$
$$E(S_{ee}^*) \xrightarrow{P} \sigma^2 + 2(1-\lambda)\beta^\top E(S_{xx})\beta,$$

as $N_M \to \infty$, where $j_i \in D_2$ is linked to $i \in D_1$, and $(x_i - x_{j_i}) = 0$ with probability λ_i. The face-value estimator of σ^2 has, therefore, an upwards bias asymptotically, which is bounded by the overall false linkage rate $1 - \lambda$, and can be adjusted accordingly.

2.6 | Asymptotic bias when using the ELE model

The ELE model treats incomplete match space as if it were complete. To examine the resulting bias, consider $\widehat{\beta}_A = (X_{D_1^*}^\top P(\lambda) X_{D_1^*})^{-1} X_{D_1^*}^\top y^*$, where

$$P(\lambda) = \lambda I_{N^* \times N^*} + \lambda_{N^*} (\mathbf{11}^\top - I)_{N^* \times N^*}, \quad \lambda_{N^*} = \frac{1-\lambda}{N^* - 1},$$
$$X_{D_1^*}^\top P(\lambda) X_{D_1^*} = G + H, \quad G = \lambda X_{D_1^*}^\top X_{D_1^*}, \quad H = \lambda_{N^*} N^* (N^* \bar{x}\bar{x}^\top - \frac{1}{N^*} X_{D_1^*}^\top X_{D_1^*}).$$

An estimate of the *overall* true match rate among the links is used as $\widehat{\lambda}$. By a Lemma due to Miller (1981): $(G + H)^{-1} = G^{-1} + (1 + g)^{-1} G^{-1} H G^{-1}$, where $g = \mathrm{tr}(HG^{-1})$, we can write

$$\widehat{\beta}_A(\lambda) = \frac{1}{\lambda} \widehat{\beta}^* - \frac{\lambda_{N^*} N}{\lambda^2 (1+g)} (X_{D_1^*}^\top X_{D_1^*})^{-1} (N^* \bar{x}\bar{x}^\top - \frac{1}{N^*} X_{D_1^*}^\top X_{D_1^*}) \widehat{\beta}^*.$$

Let $\bar{x}^\top (\frac{1}{N} X_{D_1^*}^\top X_{D_1^*})^{-1} \bar{x} \xrightarrow{P} \kappa_x$, as $N^* \to \infty$, we have

$$g = \mathrm{tr}(\frac{\lambda_{N^*} N^*}{\lambda} (N^* \bar{x}\bar{x}^\top - \frac{1}{N^*} X_{D_1^*}^\top X_{D_1^*})(X_{D_1^*}^\top X_{D_1^*})^{-1})$$
$$= \frac{\lambda_{N^*} N^*}{\lambda} (\bar{x}^\top (\frac{1}{N^*} X_{D_1^*}^\top X_{D_1^*})^{-1} \bar{x} - \frac{p}{N^*}) \quad \xrightarrow{P} \frac{1-\lambda}{\lambda} \kappa_x.$$

Let $(\frac{1}{N^*} X_{D_1^*}^\top X_{D_1^*})^{-1} \bar{x}\bar{x}^\top \to \zeta$, as $N^* \to \infty$. Provided consistent Pseudo-OLS, we have

$$\widehat{\beta}_A - \widehat{\beta}_P \xrightarrow{P} \frac{1-\lambda}{\lambda} \zeta\beta - \frac{1-\lambda}{\lambda(\kappa_x + (1-\kappa_x)\lambda)} \zeta(\beta + E(\widehat{\beta}^* - \beta)),$$

which is the asymptotic bias of $\widehat{\beta}_A$. In cases $\kappa_x \approx 1$ and $\lambda \approx 1$, the asymptotic bias is of the magnitude $(1 - \lambda)\zeta E(\widehat{\beta}^* - \beta)$, which is bounded by the false linkage rate $1 - \lambda$. Then, direct application of the ELE-model estimator can nevertheless remove almost all the bias of the face-value OLS.

Example 2 Consider $y_i = \alpha + \beta x_i + \epsilon_i$. Let $\sum_{i \in D_1^*} x_i^* / N^* \xrightarrow{P} \mu_x$ and $\sum_{i \in D_1^*} (x_i^*)^2 / N^* \xrightarrow{P} \tau_x$. We have

$$\kappa_x = \begin{bmatrix} 1 & \mu_x \end{bmatrix} \begin{bmatrix} 1 & \mu_x \\ \mu_x & \tau_x \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ \mu_x \end{bmatrix} = \frac{\tau_x - \mu_x^2}{\tau_x - \mu_x^2} = 1.$$

It follows that the asymptotic bias of $\widehat{\beta}_A$ based on the ELE model is given by

$$E(\widehat{\beta}_A - \beta) = -\frac{1-\lambda}{\lambda}\zeta E(\widehat{\beta}^* - \beta) = -\frac{1-\lambda}{\lambda}\begin{bmatrix} [1 \; \mu_x]E(\widehat{\beta}^* - \beta) \\ 0 \end{bmatrix},$$
$$\zeta = \begin{bmatrix} 1 & \mu_x \\ \mu_x & \tau_x \end{bmatrix}^{-1} \begin{bmatrix} 1 & \mu_x \\ \mu_x & \mu_x^2 \end{bmatrix} = \begin{bmatrix} 1 & \mu_x \\ 0 & 0 \end{bmatrix}.$$

In other words, the slope estimator is unbiased asymptotically, as $N^* \to \infty$; the bias of the intercept estimator is negligible as well, for example, it is about 2% of the bias of the face-value OLS if the overall false linkage rate is 2%, despite heterogeneous linkage errors. This is, thus, a favourable setting, under which the estimator can be robust against departures from ELE model assumptions.

3 | A DIAGNOSTIC TEST FOR NILE

In one form or another, assumptions of non-informative linkage errors are required in all the existing least squares methods. For $\widehat{\beta}_G$ and $\widehat{\beta}_P$ developed above, it is natural to ask if one can test whether the NILE assumption is acceptable in a given application. Provided both the estimators are consistent, we have $\widehat{\beta}_G - \widehat{\beta}_P \xrightarrow{P} 0$, as $N_G^* \to \infty$, which suggests the following diagnostic test statistic

$$t = (\widehat{\beta}_G - \widehat{\beta}_P)^\top V(\widehat{\beta}_G - \widehat{\beta}_P)^{-1}(\widehat{\beta}_G - \widehat{\beta}_P) \sim \chi_p^2 \tag{9}$$

for H_0: NILE (g1) and (p1) vs. H_1: not both (g1) and (p1). Provided asymptotic normal distribution of $\widehat{\beta}_G - \widehat{\beta}_P$, as $N_M = |M| \to \infty$, t follows the χ_p^2-distribution. The test (9) bears some resemblance to that of Hausman (1978). However, neither $\widehat{\beta}_G$ nor $\widehat{\beta}_P$ is consistent under H_1, and neither of them is fully efficient under H_0. Thus, the power of the test can be limited compared to that of Hausman (1978), and we need to derive the variance $V(\widehat{\beta}_G - \widehat{\beta}_P)$ directly.

Let D_G^* and D_P^* be the set of linked entities from D_1 under Gold and sub-Gold linkage, respectively. The variance $V(\widehat{\beta}_P)$ is given by (8) on replacing D_1^* with D_P^*, whereas $V(\widehat{\beta}_G) = (X_{D_G^*}^\top X_{D_G^*})^{-1}\sigma^2$. As shown in Appendix C, the covariance $Cov(\widehat{\beta}_G, \widehat{\beta}_P)$ can be given by

$$\begin{aligned} Cov(\widehat{\beta}_G, \widehat{\beta}_P) &\approx \frac{\sigma^2}{\lambda N_P^*} H_G(\bar{x}_G \bar{x}_G^\top + S_G^2) H_P + (1 - \frac{1}{\lambda})\frac{\sigma^2}{N_P^*} H_G \bar{x}_G \bar{x}_P^\top H_P \\ &= \frac{\sigma^2}{\lambda N_P^*} H_P - (\frac{1}{\lambda} - 1)\frac{\sigma^2}{N_P^*} H_G \bar{x}_G \bar{x}_P^\top H_P, \end{aligned}$$

where $H_G = (\frac{1}{N_G^*}\sum_{i \in D_G^*} x_i x_i^\top)^{-1}$ and $\bar{x}_G = \frac{1}{N_G^*}\sum_{i \in D_G^*} x_i$, and $H_P = (\frac{1}{N_P^*}\sum_{i \in D_P^*} x_i x_i^\top)^{-1}$ and $\bar{x}_P = \frac{1}{N_P^*}\sum_{i \in D_P^*} x_i$. Notice that, in case $\lambda \approx 1$, the covariance is dominated by the first term, and the difference between the first terms of (10) and (8) is positive definite since $1/\lambda > 1$. Moreover, λN_P^* is the asymptotic expectation of the number of true matches by sub-Gold linkage, which can easily be larger than N_G^* unless all the additional links are false. One may, therefore, expect positive definite $V(\widehat{\beta}_G) - Cov(\widehat{\beta}_G, \widehat{\beta}_P)$, since $H_P - H_G \xrightarrow{P} 0$ provided the consistency conditions for $\widehat{\beta}_G$ and $\widehat{\beta}_P$.

4 | AN APPLICATION TO INCOME DATA

The data of this application refer to administrative tax registers of income declarations in 2014 and 2015. The linkage procedure aims to connect incomes in the two years related to the same individuals. The linkage key variables are generally of good quality, though in some cases they can be missing or affected by errors. The linkage is carried out at the Italian National Institute of Statistics, and the false linkage rate is assessed to be between 1.18% and 3.76%. No information about the linkage errors at the individual level are available to us. We consider a simple linear regression model, where the income in 2014 is treated as x and that in 2015 as y. The analysis here is concerned with the data from a small locality, where there are 791 individuals in the tax register in 2014 and 771 in 2015. The linked set contains 711 individuals. A scatter plot of the associated $(x, y)_{M^*}$ is given in Figure 2. The application illustrates an advantage of using administrative data, which allows one to carry out analysis at a detailed level that cannot be supported by sample surveys otherwise.

For this linkage data set, we calculate the face-value OLS $\widehat{\beta}^*$, the estimators $\widehat{\beta}_{LL}$ and $\widehat{\beta}_A$ under the ELEmodel, as well as the Pseudo-OLS $\widehat{\beta}_P$ that allows for heterogeneous linkage errors and incomplete match space. Without information about the false linkage probabilities of the individual links, we cannot further select a Gold linkage set D_G^*, or implement the diagnostic test (9). The Gold linkage OLS $\widehat{\beta}_G$ and the diagnostic test (9) will be investigated in a simulation study in Section 5.

Table 1 shows the estimated regression coefficients and their associated confidence intervals. The face-value OLS suggests that the regression model can explain most of the variation in the dependent variable ($R^2 = 0.958$). In particular, the relative standard error of the slope

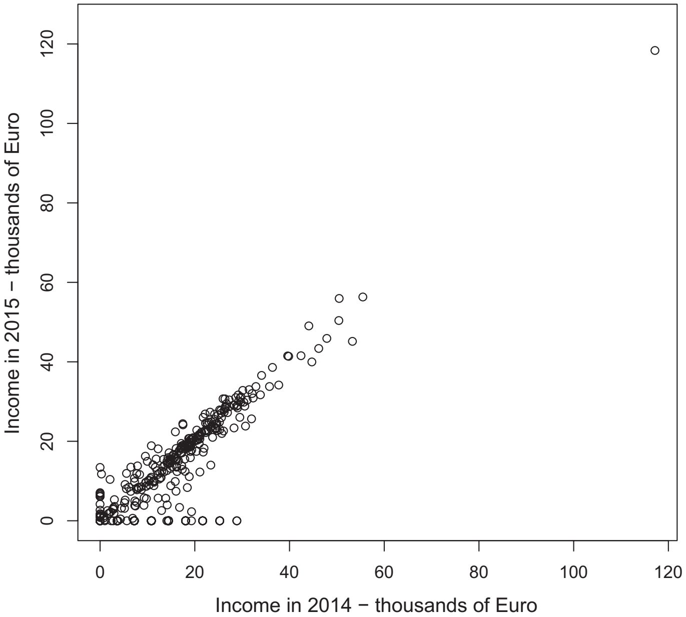


FIGURE 2 Scatter plot of linked income data in the application

TABLE 1 Estimates of year-on-year income intercept and slope, with associated confidence intervals

Estimator	Intercept	Confidence interval	Slope	Confidence interval
False linkage rate fixed at 1.18%				
$\widehat{\beta}^*$	90.644	[−114.217, 295.505]	0.983	[0.968, 0.998]
$\widehat{\beta}_{LL}$	7.191	[−242.640, 257.023]	0.994	[0.961, 1.028]
$\widehat{\beta}_A$	52.242	[−139.454, 243.938]	0.983	[0.964, 1.002]
$\widehat{\beta}_P$	7.310	[−129.794, 144.414]	0.994	[0.984, 1.005]
False linkage rate fixed at 3.76%				
$\widehat{\beta}^*$	90.644	[−114.217, 295.505]	0.983	[0.968, 0.998]
$\widehat{\beta}_{LL}$	−182.411	[−598.275, 233.451]	1.021	[0.960, 1.082]
$\widehat{\beta}_A$	−125.782	[−407.104, 155.541]	1.007	[0.976, 1.037]
$\widehat{\beta}_P$	−182.012	[−330.779, −33.246]	1.021	[1.001, 1.032]

estimator is only 0.007, which is of the same magnitude as the aforementioned false linkage rates. It follows that the bias due to the false links is not a negligible source of error, compared to the variance of the slope estimator, so that appropriate adjustment of the linkage errors is important in this case.

Fixing the overall false linkage rate $1 - \lambda$ either at 1.18% or 3.76%, the other estimates and their associated confidence intervals are given in Table 1. It can be seen that $\widehat{\beta}_A$ deviates least from the face-value OLS, for both values of λ; whereas $\widehat{\beta}_{LL}$ and $\widehat{\beta}_P$ are close to each other. However, the Pseudo OLS $\widehat{\beta}_P$ is apparently much more efficient compared to the ELE-model estimators. For example, at $1 - \lambda = 1.18\%$, the width of the confidence interval is 0.067 for the slope estimator by $\widehat{\beta}_{LL}$, whereas it is 0.021 by $\widehat{\beta}_P$, according to which the variance ratio between the two is only about 10%. The efficiency gain is somewhat greater at $1{-}\lambda = 3.76\%$.

A reason that the Pseudo-OLS can be more efficient than the ELE-model estimators is that the linkage error adjustment affects only the linked sample covariance S_{xy^*}, but not the marginal sample quantities such as the means of x and y or the matrix $X_{D_1^*}^\top X_{D_1^*}$. Of course, there is the possibility that the comparison here may be affected by the quality of variance estimation, so that the relative efficiency is not accurately assessed. We shall examine this point in the simulation study in Section 5.

The variance formula (8) allows one to incorporate the estimation uncertainty in $\widehat{\lambda}$, which is not available to the existing ELE-model estimators in closed-form expression. Since we are not provided an estimate of $V(\widehat{\lambda})$, we proceed in a practical manner as follows. Treating the reported range of false linkage rate as if it were a 95% normality-based confidence interval for $1 - \lambda$, we obtain the centre point $1 - \widehat{\lambda} = 2.47\%$ as an estimate of $1 - \lambda$, and we use the quarter length 0.645% as an estimate of $SE(\widehat{\lambda})$. Applying $\widehat{\beta}_P$ with this $\widehat{\lambda}$ and its associated estimate of $V(\widehat{\lambda})$, we obtain the regression coefficient estimates -86.099 and 1.008 for the intercept and slope, respectively, with associated confidence interval [−258.478, 86.279] for the intercept and [0.991, 1.024] for the slope. As can be expected, the point estimates are between the corresponding ones reported in Table 1. The width of the confidence interval for the slope is now 0.033, compared to 0.031 when $1{-}\lambda$ is fixed at 3.76% and 0.021 when $1{-}\lambda$ is fixed at 1.18%. Thus, it would be misleading if the inference does not take into account the uncertainty due to the estimation of λ. This is another advantage of the Pseudo-OLS method.

5 | A SIMULATION STUDY

We have four main objectives for this simulation study. First, we would like to be reassured that the apparent efficiency gains of the proposed Pseudo-OLS is not misleading. Second, a related question is the quality of associated variance estimation. Third, since one does not know to what extent the assumption of exchangeable linkage errors is violated in the application, confirmation can be obtained by simulation that the Pseudo-OLS estimator does hold in the presence of heterogeneous linkage errors. Four, we would like to investigate the performance of the diagnostic test for the NILE assumption. To the end of these objectives, we devise three scenarios below in Section 5.1.

5.1 | Set-up

5.1.1 | Scenario I: Real-life linkage and regression data

This scenario addresses all the four objectives.

The ESSnet-DI is a Eurostat project on data integration from 2009 to 2011. We use the data disseminated by Essnet DI – McLeod, Heasman and Forbes (2011), which are freely available online. The data set comprises over 26,000 individuals. It contains synthetic linkage key variables (names, dates of birth, addresses) for each individual. The key variables are distorted by missing values and typos in several different ways, which imitate real-life errors in these variables that can cause potential linkage errors. One can observe the true linkage errors by comparing the links with the true matches that are known.

For real-life regression data, we attach anonymised income data to each individual in the ESSnet-DI population, which are drawn randomly and with replacement from the linked tax data, but without being limited to the locality (in Section 4) with only 711 linked records. A scatter plot of the synthetic population income data is given in Figure 3. It can be seen that a simple linear regression model remains plausible for the simulated population values. However, there are now clearly outliers to the regression model, drawn from outside the data in the application (Figure 2). We do not remove the outliers, since it would be interesting to explore how they might affect the results.

To simulate repeated linkage and regression analysis, each time we draw first a sample of 1000 individuals from this fixed synthetic population. We then break up the sample into two separate sets D_1 and D_2, where D_1 is selected from the 1000 individuals by Bernoulli sampling with probability $\pi_1 = 0.93$, and D_2 by separate Bernoulli sampling with probability $\pi_2 = 0.92$. This creates an incomplete match space, where the expected number of matched individuals between D_1 and D_2 is $1000\pi_1\pi_2 \approx 856$.

Using a chosen set of key variables, probabilistic linkage by the approach of Fellegi and Sunter (1969) is implemented using the software Relais (2015). Over 100 simulations, the average match rate N^*_{MM}/N_M is 83.3% and the false linkage rate $1 - N^*_{MM}/N^*$ is 2.016%, that is, the sub-Gold linkage setting. For Gold linkage, we use a different set of key variables with fewer errors. Over 100 simulations, the average match rate is reduced to about 50%, while the false linkage is reduced to 0.046%. The linkage errors are heterogeneous across the different individuals.

We apply $\hat{\beta}_G$ by (5) to each Gold linkage set. For each sub-Gold linkage set, we obtain $\hat{\beta}_P$ by (6), as well as the ELE-model estimators $\hat{\beta}_{LL}$ and $\hat{\beta}_A$. For these adjustments we use the true overall false linkage rate λ in each linked set. We do not simulate additional estimation of λ, as it is not in the focus of this paper and it would affect all the adjustment methods equally. Finally, we apply the diagnostic test (9) based on $\hat{\beta}_G$ and $\hat{\beta}_P$.

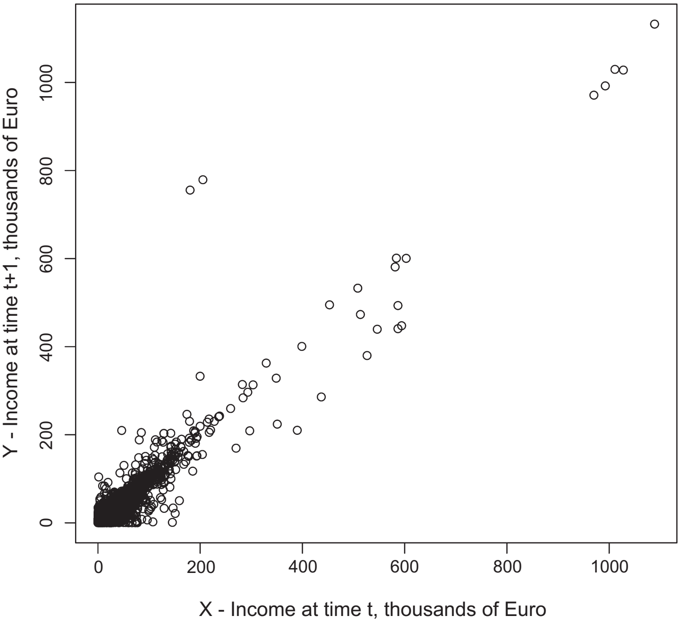


FIGURE 3 Scatter plot of synthetic income data in the ESSnet-DI population

5.1.2 | Scenario II: Real-life linkage data, artificial regression data

We expect Scenario-I can help us to better understand the application results in Section 4. Insofar as the fixed synthetic population of income data may have certain peculiar features that complicate the interpretation, we generate additional artificial regression data, reusing the simple linear regression setting of Chambers (2009), where

$$y_i = 1 + 5x_i + \epsilon_i, \quad x_i \sim \text{Uniform}(0, 1) \text{ and } \epsilon_i \sim N(0, 1).$$

Since the linear regression model holds, while the linkage errors remain uncontrolled and realistic, in Scenario-II we are able to isolate the effects of linkage errors on the estimators. The simulation of repeated linkage and regression analysis is the same as under Scenario-I, except that for each sample of 1000 individuals, we now simulate (x_i, y_i), for $i = 1, ..., 1000$, independently according to the specific regression model above. Regression analysis is then based on these (x,y)-values instead of the real-life income data.

5.1.3 | Scenario III: Artificial linkage and regression data

To confirm that $\widehat{\beta}_G$ and $\widehat{\beta}_P$ can deal with heterogeneous linkage errors under the NILE assumption, we simulate artificial linkage data by reusing the setting of Chambers (2009). For each sample of 1000 individuals, we first simulate artificial (x,y)-values as in Scenario-II. Next, the 1000 individuals are randomly divided into three blocks. The first block contains 75% of the individuals, where $\lambda_i \equiv 1$, so that these can be linked perfectly. The second block contains 15% individuals, where $\lambda_i \equiv 0.95$, so that the linkage results would be fairly good for them. The third block contains the remaining 10% individuals, where $\lambda_i \equiv 0.75$, and the linkage results would be rather poor for them. Moreover, we do not simulate subsampling of D_1 and D_2, so that we have complete match space by construction. The linked set can now be simulated directly, without actually implementing any linkage procedure. Had we broken up the sample into D_1 and D_2, dividing the 1000 records in three blocks, and linked every record in D_1 to one in D_2 from the same block, the linkage errors would have been on expectation the same as we have just specified.

This yields an overall false linkage rate that equals to 0.9675, which is quite close to that in Scenario-I (and II). Given each simulated linkage set, we calculate $\widehat{\beta}_P$ using a single adjustment factor $\lambda = 0.9675$, and the ELE-model estimators given block-diagonal P-matrix with known λ-values. We can see how well $\widehat{\beta}_P$ handles heterogenous linkage errors by comparing it to the benchmark ELE-model estimators. Finally, we simply calculate $\widehat{\beta}_G$ based on the first-block of links.

5.2 | Results of regression coefficient estimators

Figure 4 shows the Percentage Relative Errors (PREs) of the different regression coefficient estimates. For each linked set, the 'error' of an estimate is calculated as its difference to the corresponding true OLS estimate $\tilde{\beta}$, based on the matched individuals D_M as when linkage is unnecessary. Over

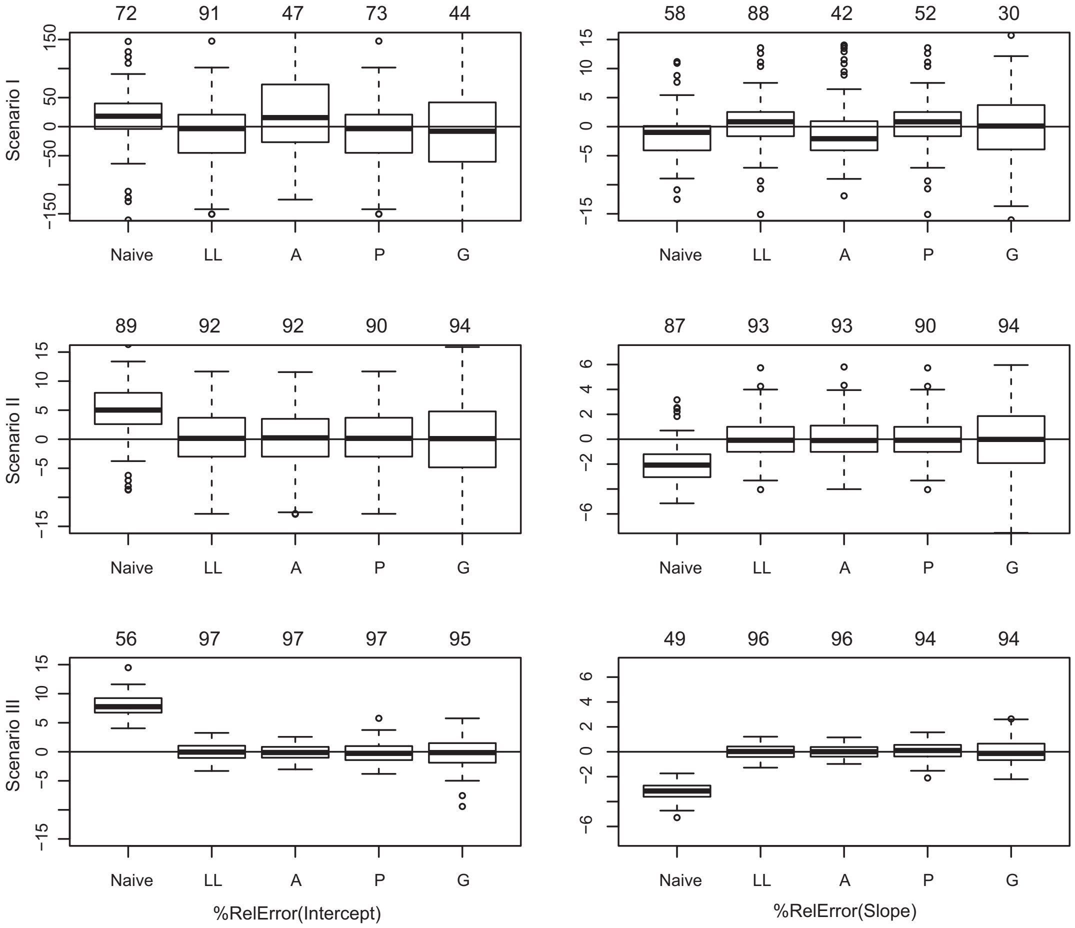


FIGURE 4 Boxplot of PREs of intercept and slope estimates: Scenario-I (top), II (middle), III (bottom); estimator $\widehat{\beta}_{LL}$ (LL), $\widehat{\beta}_A$ (A), $\widehat{\beta}_P$ (P) and $\widehat{\beta}_G$ (G); coverage of 95% confidence interval (over top margin)

the top margin of each box-plot, we report the actual coverage rates of the nominal 95% confidence intervals using the associated variance estimators. Table 2 provides the empirical standard error (SE) of each estimator over the 100 simulations, and the corresponding average of the 100 SE estimates.

Consider the results under Scenario-I, which are immediately relevant to those in Section 4. First, as expected, the presence of false links weakens the observed correlation between x_i and y_i^*. Hence, the face-value estimate of the slope is negatively biased when the true slope is positive, and the intercept estimate is biased in the opposite direction. It can be seen in Figure 4 that all the adjusted estimators are less biased than the face-value OLS, where $\hat{\beta}_{LL}$ and $\hat{\beta}_P$ have the most similar expectations, which is compatible with the application results in Table 1, where these two estimators are closest to each other. Moreover, it illustrates that in case of heterogeneous but low false linkage probabilities, the ELE-model estimators can nevertheless remove most of the bias, as discussed in Section 2.6.

Next, according to the SEs in Table 2, the Pseudo-OLS is the most efficient of all the linkage-data estimators, including the face-value OLS. The relative efficiency to the ELE-model estimators is comparable to that estimated in Table 1. This suggests that the gains are genuine in the application. Since $\hat{\beta}_P$ is calculated using λ instead of its estimate here, the efficiency gains against $\hat{\beta}_G$ are somewhat overstated. Nevertheless, generally one may expect $\hat{\beta}_P$ to be more efficient than $\hat{\beta}_G$, as long as the effective sample size N_{MM}^* is much larger based on sub-Gold linkage (e.g. with about 2% false linkage rate here) than based on Gold linkage (e.g. with about 50% missing match rate here).

Meanwhile, the means of the SE estimators (Table 2) over the 100 simulations show that all the variances are over-estimated considerably, including the true OLS, and the coverage of the 95% confidence intervals are very erratic. This is mainly caused by the regression model outliers in this case, as noticed earlier for Figure 3. Thus, these results serve well as a reminder that, in linkage-data regression, one must not forget about the problems that can also cause troubles in the absence of linkage errors. Notice that variance over-estimation is not a problem for the application results in Table 1, where critical outliers are absent from the linked data set (Figure 2).

When it comes to Scenario-II, we can see in Figure 4 that all the adjusted estimators are nearly unbiased, as can be expected given the results under Scenario-I. The Pseudo-OLS remains the most

TABLE 2 Results for variance estimation over 100 simulations

Scenario		True	Naïve	$\hat{\beta}_{LL}$	$\hat{\beta}_A$	$\hat{\beta}_P$	$\hat{\beta}_G$
Intercept							
I	Standard error	386.1	457.1	1222.7	575.8	388.9	545.1
	SE estimator	2431.7	2604.5	2645.2	1256.1	2645.2	3233.6
II	Standard error	0.069	0.077	0.079	0.079	0.075	0.098
	SE estimator	0.078	0.086	0.086	0.087	0.086	0.098
III	Standard error	0.043	0.042	0.043	0.044	0.043	0.051
	SE estimator	0.045	0.048	0.048	0.047	0.046	0.052
Slope							
I	Standard error	0.012	0.015	0.052	0.022	0.013	0.018
	SE estimator	0.113	0.118	0.120	0.057	0.120	0.149
II	Standard error	0.119	0.134	0.138	0.138	0.131	0.171
	SE estimator	0.131	0.149	0.150	0.151	0.150	0.160
III	Standard error	0.075	0.074	0.075	0.074	0.077	0.081
	SE estimator	0.078	0.083	0.084	0.083	0.080	0.089

efficient linkage-data method. The results of variance estimation appear acceptable for all the estimators, now that outlier-contaminated income data are replaced by true regression data. While there still exists some slight over-estimation of the variance, it is not related to the adjustment methods, because the amount of over-estimation for them is comparable to that for the true OLS. The coverage of the confidence interval derived from the face-value OLS is improved compared to that in Scenario-I, because its bias is relatively small here. Nevertheless, bias adjustment is preferable.

The ELE model assumptions of $\widehat{\beta}_{LL}$ and $\widehat{\beta}_{A}$ are fully satisfied in Scenario-III. Likewise for $\widehat{\beta}_{G}$ under the NILE assumption. Despite $\widehat{\beta}_{P}$ uses only an overall false linkage rate, Figure 4 shows clearly that it is as effective as the benchmark estimators at reducing the bias due to the linkage errors. This confirms that the Pseudo-OLS can accommodate heterogeneous linkage errors in a simple manner, provided the NILE assumption is satisfied. The Pseudo-OLS is no longer the most efficient method here, which is not surprising given that the assumptions of the other estimators are exactly satisfied. The principal advantages of the Pseudo-OLS lies in real-life situations, where the match space is incomplete and the secondary analyst has no detailed knowledge of the record linkage procedure, such as the three blocks of linkage errors in this case. The results of variance estimation are acceptable for all the estimators. Due to increased bias relative to its variance, the face-value OLS again leads to low coverage here. The coverages rates derived from ($\widehat{\beta}_{LL}$, $\widehat{\beta}_{A}$, $\widehat{\beta}_{P}$, $\widehat{\beta}_{G}$) deviate from the nominal 95% level by one or two percentage points in Figure 4. It is reassuring to notice that this is simply due to the Monte Carlo error of the 100 simulations, because all the coverage rates converge to 95% as we increase the number of simulations to 1000, now that the assumptions of the benchmark estimators are satisfied here.

5.3 | Results of diagnostic test

The results of the diagnostic test for the NILE assumption are given in Figure 5. Under each scenario, the histogram of the test statistic values over 100 simulations are compared to the χ^2 density function with 2 degrees of freedom, which is the distribution under the null hypothesis. At the 5% significance level, the rejection rate over the 100 simulations is 0.63 under Scenario-I, 0.06 under Scenario-II and 0.02 under Scenario-III.

Take first Scenario-III, where the set-up satisfies both the NILE assumptions and the regression model, the histogram of the test statistic values agrees reasonably well with its theoretical null distribution. Provided the relevant NILE assumptions, the higher missing-match rate of Gold linkage and the heterogenous linkage errors of sub-Gold linkage on expectation do not lead to unbalanced selection of the linked entities under either. Over the 100 simulations, the rejection rate of the diagnostic test at the 5%-level is 0.02, which appears to agree with the actual performance of $\widehat{\beta}_{P}$ and $\widehat{\beta}_{G}$.

Next, the set-up of Scenario-II satisfies the regression model assumptions, but it does not necessarily fulfil the NILE assumption *a priori*, since the errors of the key variables had been generated in ways which imitate real-life idiosyncrasies that are beyond our control. However, over the 100 simulations, the empirical SE of $\widehat{\beta}_{G} - \widehat{\beta}_{P}$ are 0.069 and 0.130 for the difference of intercept and slope, respectively, whereas the average of the corresponding SE estimates are 0.066 and 0.116. The histogram of the test statistic values agrees fairly well with its theoretical null distribution. The rejection rate of the 5%-level test is 0.06, which again seems reasonable in light of the actual performance of $\widehat{\beta}_{P}$ and $\widehat{\beta}_{G}$ in Figure 4. These are evidences suggesting that the relevant NILE assumptions can be met at least approximately in many practical situations.

Meanwhile, the test performance deteriorates under Scenario-I with real-life data for regression. For instance, the histogram of the test statistic values does not agree at all with the theoretical null distribution. The rejection rate of the 5%-level test is 0.63, which is unnecessarily high in light of the

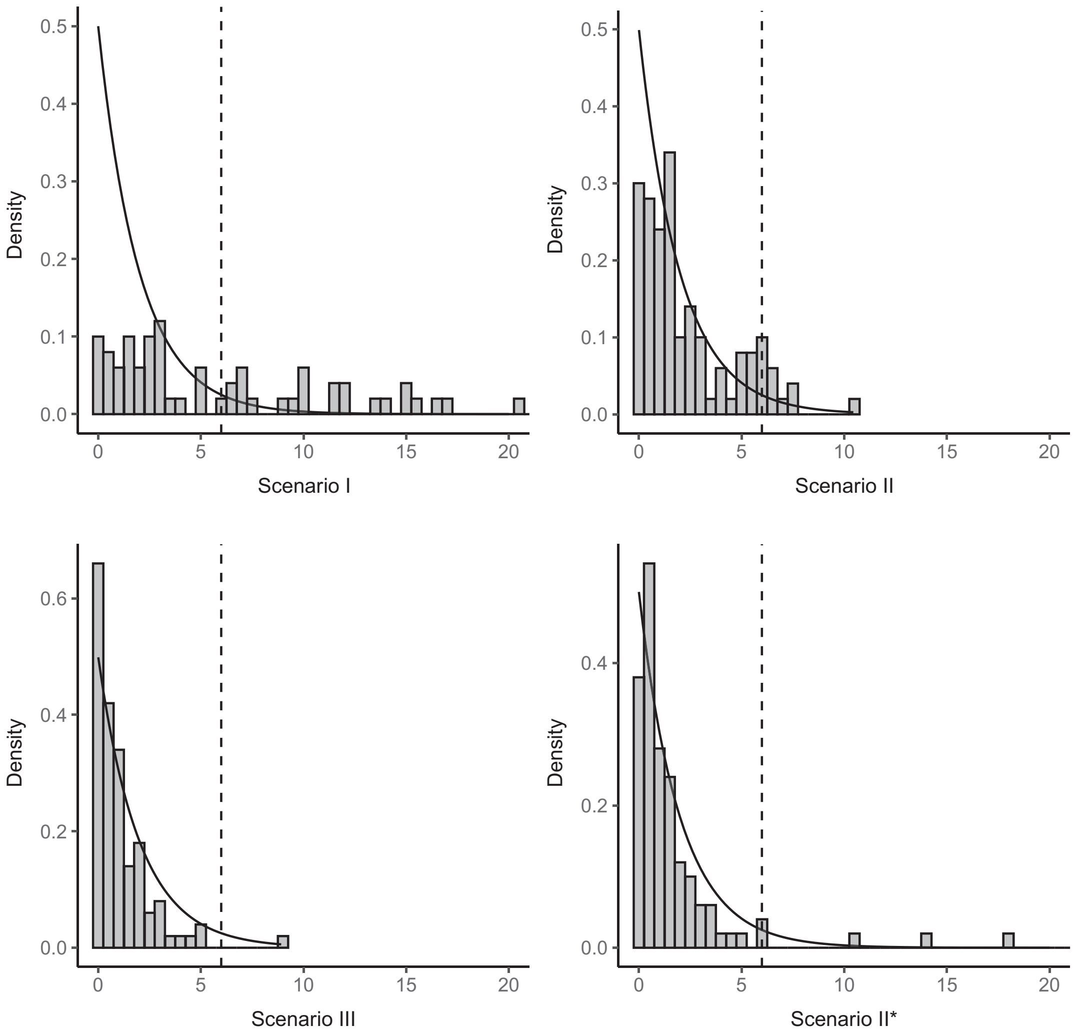


FIGURE 5 Diagnostic test for NILE assumption under Scenario-I to III: χ^2_2 density function (solid) with 95^{th} percentile (vertical dashed), histogram of observed test values over 100 simulations. Additional Scenario-II$*$ with rejection ratio 0.04 over 100 simulations

bias reduction that can be achieved by $\widehat{\beta}_P$ and $\widehat{\beta}_G$ here. The imbalance of regression outliers between the two linkage sets causes severe under-estimation of $V(\widehat{\beta}_G - \widehat{\beta}_P)$. For example, the empirical SE is 2452.5 for the difference in intercept estimates and 0.113 for the slope difference, but the average of the corresponding SE estimates is only 446.7 and 0.015, respectively. The severely under-estimated denominator of the test statistic (9) leads then to the high rejection rate over the simulations.

For confirmation we carried out additional simulations, where we simulated the 3-block ELE linkage errors, while retaining the real-life income data for regression. The results are shown in Figure 5, designated as Scenario-II*, which are similar to those under Scenario-II and III. The empirical SE is 1380.8 for the intercept difference and 0.061 for the slope, while the average of SE estimates is 1721.1 and 0.072, respectively, despite the presence of regression outliers. The rejection rate of the 5%-level test is now 0.04, which would be more helpful in practice. The cause of these results lies in the different set-ups of Scenario-I and II*. Although regression outliers are present in both cases, the linkage errors are randomly 'assigned' to the sample units under Scenario-II*, such that they may affect 'evenly' $\widehat{\beta}_P$ and $\widehat{\beta}_G$ over repeated simulations. Under Scenario-I, however, the linkage key variables are fixed for each individual, such that, for example, regression outliers affect only $\widehat{\beta}_P$ but not $\widehat{\beta}_G$, provided the outliers in the population happen to be associated only with sub-Gold linkage individuals but none of the Gold linkage individuals. Such peculiarities in the *fixed* population of regression and key variables can affect the simulation results unexpectedly.

Finally, taking altogether the test results from the simulation study here, it would seem reasonable that one should not interpret the p-value of the diagnostic test too stringently in practice, for example, despite the p-value is only 0.05 or even slightly lower in a given situation, the estimators assuming non-informative linkage errors are still likely be very helpful.

6 | CONCLUDING REMARKS

Heterogenous linkage errors and incomplete match space are likely to prevail in most applications of record linkage. We propose a practical approach to linkage-data regression for secondary analysis, which can accommodate both in a simple manner, provided suitable NILE assumptions of the linkage errors. Application and simulation suggest that the relevant assumptions can be met at least approximately in many situations. In the simulation studies where the match space is incomplete, the proposed Pseudo-OLS method is more efficient than the existing adjustment methods that operate under the approximate assumption of complete match space. Moreover, we construct for the first time an accompanying diagnostic test for the NILE assumptions, which can provide helpful guidance in practice. Regarding future development, we believe additional research is needed for robust variance estimation, which can better cope with heterogeneous regression errors and potential outliers. As another current research topic we are developing an extension of our approach to categorical linkage-data analysis.

REFERENCES

Asher, J., Banks, D. & Scheuren, F.J. (2008) *Statistical methods for human rights*. New York: Springer.

Chambers, R. (2009) Regression analysis of probability-linked data. *Official Statistics Research Series*, Vol. 4. Statistics New Zealand.

Chambers, R.C. & Kim, G. (2015) Secondary analysis of linked data. In: Harron, K., Goldstein H. & Dibben, C. (Eds.) *Methodological developments in data linkage*, Chapter 5. Hoboken: John Wiley & Sons. 83–108.

Chambers, R. & Diniz da Silva, A.D. (2020) Improved secondary analysis of linked data: A framework and an illustration. *Journal of the Royal Statistical Society: Series A*, 183, 37–59. 10.1111/rssa.12477.

Chipperfield, J.O. & Chambers, R.C. (2015) Using bootstrap to account for linkage errors when analysing probabilistically linked categorical data. *Journal of Official Statistics*, 31, 397–414.

Chipperfield, J.O., Bishop, G.R. & Campell, P. (2011) Maximum likelihood estimation for contingency tables and logistic regression with incorrectly linked data. *Survey Methodology*, 37, 13–24.

Christen, P. (2012) A survey of indexing techniques for scalable record linkage and deduplication. *ISEE Transactions on Knowledge and Data Engineering*, 24, 1537–1555.

Copas, J.B. & Hilton, F.J. (1990) Record linkage: Statistical models for matching computer records. *Journal of the Royal Statistical Society, Series A,*, 153, 287–320.

Creel, S., Spong, G., Sands, J.L., Rotella, J., Zeigle, J., Joe, L., & Smith, D. (2003) Population size estimation in Yellowstone wolves with erro-prone noninvasive microsatellite genotypes. *Molecular Ecology*, 12(7), 2003–2009.

Enamorado, T., Fifield, B. & Imai, K. (2019) Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review*, 113, 353–371.

Essnet DI – McLeod, Heasman and Forbes. (2011) Simulated data for the on the job training, available at https://ec.europa.eu/eurostat/cros/content/job-training_en.

Fellegi, I.P. & Sunter, A.B. (1969) A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183–1210.

Goldstein, H., Harron, K. & Wade, A. (2012) The analysis of record-linked data using multiple imputation with data value priors. *Statistics in Medicine*, 31, 3481–3493.

Gutman, R., Afendulis, C.C. & Zaslavsky, A.M. (2013) A Bayesian procedure for file linking to analyze end-of-life medical costs. *Journal of the American Statistical Association*, 108, 34–47.

Gutman, R., Sammartino, C.J., Green, T.C. & Montague, B.T. (2015) Error adjustments for file linking methods using encrypted unique client identifier (eUCI) with application to recently released prisoners who are HIV+. *Statistics in Medicine*, 35, 115–129.

Han, Y. & Lahiri, P. (2018) Statistical analysis with linked data. *International Statistical Review*, 87, S139–S157. 10.1111/insr.12295.

Harron, K., Gilbert, K., Cromwell, D. & van der Meulen, J. (2016) Linking data for mothers and babies in de-identified electronic health data. *PLoS ONE*, 11(10), e0164667. 10.1371/journal.pone.0164667.

Harron, K., Goldstein, H. & Dibben, C. (2015) *Methodological developments in data linkage*. Hoboken: Wiley.

Hausman, J.A. (1978) Specification tests in econometrics. *Econometrica*, 46, 1251–1271.

Van der Heijden, P.G., Cruyff, M., & Böhning, D. (2014) *Capture recapture to estimate criminal populations. Encyclopedia of criminology and criminal justice*. Berlin: Springer, pp. 267–276.

Herzog, T.N., Scheuren, F.J. & Winkler, W.E. (2007) *Data quality and record linkage techniques*. Berlin: Springer.

Hof, M.H.P. & Zwinderman, A.H. (2012) Methods for analysing data from probabilistic linkage strategies based on partially identifying variables. *Statistics in Medicine*, 31, 4231–4242.

Jaro, M.A. (1989) Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, 84, 414–420.

Kim, G. & Chambers, R.C. (2012a) Regression analysis under incomplete linkage. *Comutational Statistics and Data Analysis*, 56, 2756–2770.

Kim, G. & Chambers, R.C. (2012b) Regression analysis under probabilistic multi-linkage. *Statistica Neerlandica*, 66, 64–79.

Lahiri, P. & Larsen, M.D. (2005) Regression analysis with linked data. *Journal of the American Statistical Association*, 100, 222–230.

Link, W.A., Yoshizaki, J., Bailey, L.L. & Pollock, K.H. (2010) Uncovering a latent multinomial: Analysis of mark–recapture data with misidentification. *Biometrics*, 66(1), 178–185.

Marchant, N.G., Steorts, R.C., Kaplan, A., Rubinstein, B.I.P. & Elazar, D.N. (2019) d-blink: Distributed End-to-End Bayesian Entity Resolution. Available from: https://arxiv.org/pdf/1909.06039.pdf.

McClintock, B.T., Bailey, L.L., Dreher, B.P. & Link, W.A. (2014) Probit models for capture–recapture data subject to imperfect detection, individual heterogeneity and misidentification. *The Annals of Applied Statistics*, 8(4), 2461–2484.

Miller, K.S. (1981) On the inverse of the sum of matrices. *Mathematics Magazine*, 54, 67–72.

Neter, J., Maynes, E.S. & Ramanathan, R. (1965) The effect of mismatching on the measurement of response error. *Journal of the American Statistical Association*, 60, 1005–1027.

Abbott, O., Jones, P., & Ralphs, M. (2015) *Large-scale linkage for total populations in official statistics. Methodological Developments in Data Linkage*, pp. 170-200.

RELAIS 3.0 Users Guide. (2015). Available from: http://www.istat.it/it/strumenti/metodi-e-strumenti-it/strumenti-di-elaborazione/relais.

Rosman, D.L. (2001) The Western Australian Road Injury Database (1987–1996): Ten years of linked police, hospital and death records of road crashes and injuries. *Accident Analysis & Prevention*, 33(1), 81–88.

Sadinle, M. (2014) Detecting duplicates in a homicide registry using a Bayesian partitioning approach. *Annals of Applied Statistics*, 8, 2404–2434.

Sadinle, M. (2017) Bayesian estimation of bipartite matchings for record linkage. *Journal of the American Statistical Association*, 112, 600–612

Scheuren, F. & Winkler, W.E. (1993) Regression analysis of data files that are computer matched. *Survey Methodology*, 19, 39–58.

Scheuren, F. & Winkler, W.E. (1997) Regression analysis of data files that are computer matched – Part II. *Survey Methodology*, 23, 157–165.

Seybolt, T.B., Aronson, J.D. & Fischhoff, B. (Eds.) (2013) *Counting civilian casualties: An introduction to recording and estimating nonmilitary deaths in conflict*. Oxford: Oxford University Press.

Stoerts, R., Hall, R. & Fienberg, S. (2017) A Bayesian approach to graphical record linkage and de-duplication. *Journal of the American Statistical Association*, 111, 1660–1672

Tancredi, A. & Liseo, B. (2011) A hierarchical Bayesian approach to record linkage and population size problems. *The Annals of Applied Statistics*, 5, 1553–1585.

Tuoto, T. (2016) New proposal for linkage error estimation. *Statistical Journal of the IAOS*, 32(2), 1–8.

Wright, J.A., Barker, R.J., Schofield, M.R., Frantz, A.C., Byrom, A.E. & Gleeson, D.M. (2009) Incorporating genotype uncertainty into mark–recapture-type models for estimating abundance using DNA samples. *Biometrics*, 65(3), 833–840.

Zhang, L.-C. (2019) On secondary analysis of datasets that cannot be linked without errors. In: Zhang, L.-C. & Chambers, R. L. (Eds.) *Analysis of integrated data*. London: CRC/Chapman and Hall.

Zhang, G. & Campbell, P. (2012) Data survey: Developing the statistical longitudinal census dataset and identifying its potential uses. *Australian Economic Review*, 45, 125–133.

How to cite this article: Zhang LC, Tuoto T. Linkage-data linear regression. *J R Stat Soc Series A*. 2020;00:1–26. https://doi.org/10.1111/rssa.12630

APPENDIX A

PROOF OF PROPOSITION 2

Provided (p1), that is, (4) over D_1, we have $\sum_{i \in D_1^*} x_i x_i^\top / N^* - \sum_{i \in D_1} x_i x_i^\top / N_1 \xrightarrow{P} 0$, by the same argument as in Section 2.3, where $z_i = x_i x_i^\top$ for $i \in D_1$. Provided (p0.2) in addition, we have $\sum_{i \in D_1^*} x_i x_i^\top / N^* - \sum_{i \in D_M} x_i x_i^\top / N_M \xrightarrow{P} 0$. Likewise, $\sum_{i \in D_1^*} x_i / N^* - \sum_{i \in D_M} x_i / N_M \xrightarrow{P} 0$, and $\sum_{i \in D_1^*} y_i^* / N^* - \sum_{i \in D_M} y_i / N_M \xrightarrow{P} 0$ by (p1), that is, (4) over D_2, and (p0.3). Notice that the conditions (p0.2) and (p0.3) are needed to ensure that false links of the unmatched records do not cause asymptotic bias to the 'marginal' statistics, that is, $\sum_{i \in D_1^*} x_i x_i^\top / N^*$ and $\sum_{i \in D_1^*} x_i / N^*$ based on D_1 and $\sum_{i \in D_1^*} y_i^* / N^*$ based on D_2. Finally, provided (p1), that is, (2) over D_1^*, and (p0.1), we have, as discussed in Section 2.4,

$$\begin{aligned} \sum_{i \in D_1^*} Cov(x_i, y_i^* \mid \ell_i = 1) &= \sum_{i \in D_{MM}^*} Cov(x_i, y_i) + \sum_{i \in D_{1M}^* \setminus D_{MM}^*} 0 + \sum_{i \in D_1^* \setminus D_{1M}^*} 0 \\ &= \sum_{i \in D_M} (\ell_i a_{ii}) Cov(x_i, y_i). \end{aligned}$$

Let $z_i = Cov(x_i, y_i)$ for $i \in D_M$, which is an unknown constant associated with $i \in D_M$. Now that the inclusion probability of $i \in D_{MM}^*$ from D_M is $\lambda_i \psi_i$, (p1) entails asymptotic NILE for $\ell_i a_{ii}$ over D_M, such that $\sum_{i \in D_{MM}^*} z_i / N_{MM}^* - \sum_{i \in D_M} z_i / N_M \xrightarrow{P} 0$. Notice that each term of S_{xy^*} from D_{MM}^* is an asymptotically unbiased estimate of the corresponding $Cov(x_i, y_i)$, and each term outside of D_{MM}^* has asymptotic expectation zero, so that $\lambda^{-1} S_{xy^*} - S_{xy}(M) \xrightarrow{P} 0$ given (p2), where $S_{xy}(M)$ is the empirical covariance of (x_i, y_i) over D_M. The consistency of $\hat{\lambda}$ implies then $\hat{\beta}_P - \tilde{\beta} \xrightarrow{P} 0$.

APPENDIX B

APPROXIMATE VARIANCE $V(\bar{x}\bar{y}^* + \hat{\lambda}^{-1}S_{xy^*})$

We have $V(\bar{x}\bar{y}^*|A) = \bar{x}\bar{x}^\top\sigma^2/N^*$ and $E(\bar{x}\bar{y}^*|A) = \bar{x}\bar{x}^{*\top}\beta$, where $\bar{x}^* = \sum_{j\in D_2^*} x_j/N^* \neq \bar{x} = \sum_{i\in D_1^*} x_i/N^*$. Conditional on all the x's, we obtain

$$V(\bar{x}\bar{y}^*) = E[V(\bar{x}\bar{y}^*|A)] + V[E(\bar{x}\bar{y}^*|A)] = \bar{x}\bar{x}^\top\sigma^2/N^*.$$

Next, let $v_1 = V(\sum_{i\in D_1^*}(x_i - \bar{x})(y_i^* - \bar{y}^*))$, where

$$\begin{aligned}
v_1 &= \sum_{i\in D_1^*}(x_i-\bar{x})(x_i-\bar{x})^\top V(y_i^*-\bar{y}^*) + \sum_{i\in D_1^*}\sum_{k\neq i}(x_i-\bar{x})(x_k-\bar{x})^\top Cov(y_i^*-\bar{y}^*, y_k^*-\bar{y}^*)\\
&= \sum_{i\in D_1^*}(x_i-\bar{x})(x_i-\bar{x})^\top(1-\frac{1}{n})\sigma^2 - \sum_{i\in D_1^*}\sum_{k\neq i}(x_i-\bar{x})(x_k-\bar{x})^\top\frac{1}{n}\sigma^2\\
&= \sum_{i\in D_1^*}(x_i-\bar{x})(x_i-\bar{x})^\top\sigma^2 - \frac{\sigma^2}{n}\sum_{i\in D_1^*}(x_i-\bar{x})\sum_{k\in D_1^*}(x_k-\bar{x})^\top\\
&= N^*S_{xx}\sigma^2, \qquad \text{for} \quad S_{xx} = \frac{1}{N^*}\sum_{i\in D_1^*}(x_i-\bar{x})(x_i-\bar{x})^\top,
\end{aligned}$$

since $\sum_{k\in D_1^*}(x_k - \bar{x}) = 0$, such that $V(\hat{\lambda}^{-1}S_{xy^*}|A) = S_{xx}\sigma^2/(\hat{\lambda}^2N^*)$. Moreover,

$$E(\hat{\lambda}^{-1}S_{xy^*}|A) = \hat{\lambda}^{-1}\frac{1}{N^*}\sum_{i\in D_1^*}(x_i-\bar{x})E(y_i^*-\bar{y}^*|A) = \hat{\lambda}^{-1}S_{xx^*}\beta \approx \hat{\lambda}^{-1}\lambda S_{xx}\beta,$$

where $S_{xx^*} = \sum_{i\in D_1^*}(x_i - \bar{x})(x_{j_i} - \bar{x}^*)^\top/N^*$, and x_{j_i} is the x-vector for y_j that is linked to the record i in D_1, which is uncorrelated to x_i unless $j_i = i$. Therefore, asymptotically as $|M|\to\infty$, we have $S_{xx^*} \approx S_{xx}N_{MM}^*/N^* \approx \lambda S_{xx}$. We obtain

$$V(\hat{\lambda}^{-1}S_{xy^*}) = E(\frac{\sigma^2}{\hat{\lambda}^2N^*}S_{xx}) + V(\hat{\lambda}^{-1}\psi S_{xx}\beta) \approx \frac{\sigma^2}{\lambda^2N^*}S_{xx} + V(\hat{\lambda})S_{xx}\beta\beta^\top S_{xx}^\top.$$

Putting together $V(\bar{x}\bar{y}^*)$ and $V(\hat{\lambda}^{-1}S_{xy^*})$ from above, we have

$$\begin{aligned}
V(\bar{x}\bar{y}^* + \hat{\lambda}^{-1}S_{xy^*}) &\approx (\bar{x}\bar{x}^\top + S_{xx})\frac{\sigma^2}{N^*} + \Delta = (\frac{1}{N^*}X_{D_1^*}^\top X_{D_1^*})\frac{\sigma^2}{N^*} + \Delta,\\
\Delta &= (\frac{1}{\lambda^2} - 1)\frac{\sigma^2}{N^*}S_{xx} + V(\hat{\lambda})S_{xx}\beta\beta^\top S_{xx}^\top,\\
V(\hat{\beta}_P) &= (X_{D_1^*}^\top X_{D_1^*})^{-1}\sigma^2 + (\frac{1}{N^*}X_{D_1^*}^\top X_{D_1^*})^{-1}\Delta(\frac{1}{N^*}X_{D_1^*}^\top X_{D_1^*})^{-1}.
\end{aligned}$$

APPENDIX C

COVARIANCE OF $\hat{\beta}_G$ AND $\hat{\beta}_P$

The estimator $\hat{\beta}_G$ given by (5) can be rewritten as

$$\begin{aligned}
&\hat{\beta}_G = H_G(\bar{x}_G\bar{y}_G + \frac{1}{N_G^*}\tau_G) \qquad \bar{x}_G = \frac{1}{N_G^*}\sum_{i\in D_G^*}x_i \qquad \bar{y}_G = \frac{1}{N_G^*}\sum_{i\in D_G^*}y_i\\
&H_G = (\frac{1}{N_G^*}\sum_{i\in D_G^*}x_ix_i^\top)^{-1} \qquad \tau_G = \sum_{i\in D_G^*}(x_i-\bar{x}_G)(y_i-\bar{y}_G) = \sum_{i\in D_G^*}(x_i-\bar{x}_G)y_i
\end{aligned}$$

By definition we have $D_G^* \subset D_P^*$ and $N_G^* < N_P^*$. Let $D_A^* = D_P^* \setminus D_G^*$ consist of the remaining entities. Let $w = N_G^*/N_P^*$, and $1 - w = N_A^*/N_P^*$. We have

$$\begin{aligned}
\widehat{\beta}_P &= H_P\big(\bar{x}_P\bar{y}_P + \frac{1}{\widehat{\lambda} N_P^*}\tau_P\big) \quad \bar{x}_P = \frac{1}{N_P^*}\sum_{i\in D_P^*} x_i \quad H_P = \big(\frac{1}{N_P^*}\sum_{i\in D_P^*} x_i x_i^\top\big)^{-1}\\
\bar{y}_P &= \frac{1}{N_P^*}\sum_{i\in D_P^*} y_i^* = w\bar{y}_G + (1-w)\bar{y}_A^* \quad \bar{y}_A^* = \frac{1}{N_A^*}\sum_{i\in D_A^*} y_i^*\\
\tau_P &= \sum_{i\in D_P^*}(x_i - \bar{x}_P)(y_i^* - \bar{y}_P^*) = \sum_{i\in D_P^*}(x_i-\bar{x}_P)y_i^* = \tau'_G + \tau_A\\
\tau'_G &= \sum_{i\in D_G^*}(x_i-\bar{x}_P)y_i \quad \tau_A = \sum_{i \in D_A^*}(x_i-\bar{x}_P)y_i^*
\end{aligned}$$

Notice that $\tau_G \neq \tau'_G$ because τ_G involves $\bar{x}_G$ whereas τ'_G involves $\bar{x}_P$. Now, to obtain the covariance, we only need to take the cross terms one by one. We have

$$Cov(H_G \bar{x}_G \bar{y}_G, H_P \bar{x}_P \bar{y}_P) = wH_G\bar{x}_G V(\bar{y}_G)\bar{x}_P^\top H_P^\top = \frac{\sigma^2}{N_P^*} H_G \bar{x}_G \bar{x}_P^\top H_P$$

because $Cov(\bar{y}_G, \bar{y}_A^*) = 0$ and $H_P = H_P^\top$. Similarly, $Cov(\bar{y}_G, \tau_A) = 0$, such that

$$\begin{aligned}
&Cov\big(H_G\bar{x}_G\bar{y}_G, \frac{1}{\widehat{\lambda} N_P^*}H_P\tau_P\big) = E\big(\frac{1}{\widehat{\lambda} N_P^*}\big)Cov(H_G\bar{x}_G \bar{y}_G, H_P\tau'_G)\\
&\approx \frac{\sigma^2}{\lambda N_P^*}H_G\bar{x}_G(\bar{x}_G - \bar{x}_P)^\top H_P^\top = \frac{\sigma^2}{\lambda N_P^*}H_G\bar{x}_G\bar{x}_G^\top H_P - \frac{\sigma^2}{\lambda N_P^*}H_G \bar{x}_G\bar{x}_P^\top H_P\\
&Cov\big(H_P\bar{x}_P\bar{y}_P, \frac{1}{N_G^*}H_G\tau_G\big) = \frac{w\sigma^2}{N_G^*}H_P\bar{x}_P(\bar{x}_G - \bar{x}_G)^\top H_G^\top = 0
\end{aligned}$$

Finally, let $S_G^2 = \sum_{i\in D_G^*}(x_i - \bar{x}_G)(x_i - \bar{x}_G)^\top/N_G^*$, such that

$$\begin{aligned}
&Cov(\tau_G,\tau'_G) = \sigma^2\sum_{i\in D_G^*}(x_i-\bar{x}_G)(x_i-\bar{x}_P)^\top = \sigma^2 N_G^* S_G^2\\
&Cov\big(\frac{1}{N_G^*}H_G\tau_G, \frac{1}{\widehat{\lambda} N_P^*}H_P\tau_P\big) = E\big(\frac{1}{\widehat{\lambda} N_P^* N_G^*}\big)H_G Cov(\tau_G,\tau'_G)H_P^\top \approx \frac{\sigma^2}{\lambda N_P^*}H_G S_G^2 H_P
\end{aligned}$$

Summarising the four terms above, and noting $H_G = \bar{x}_G\bar{x}_G^\top + S_G^2$, we obtain

$$\begin{aligned}
Cov(\widehat{\beta}_G,\widehat{\beta}_P) &\approx \frac{\sigma^2}{\lambda N_P^*}H_G(\bar{x}_G\bar{x}_G^\top + S_G^2)H_P + \big(1-\frac{1}{\lambda}\big)\frac{\sigma^2}{N_P^*}H_G\bar{x}_G\bar{x}_P^\top H_P\\
&= \frac{\sigma^2}{\lambda N_P^*}H_P - \big(\frac{1}{\lambda} - 1\big)\frac{\sigma^2}{N_P^*}H_G\bar{x}_G\bar{x}_P^\top H_P
\end{aligned}$$