When adjusting for the bias due to linkage errors: A sensitivity analysis

Loredana Di Consiglio and Tiziana Tuoto*
Italian National Statistical Institute, Rome, Italy

Abstract. Linkage of different data sources is an intermediate step in many statistical processes. When dealing with data resulting from a record linkage process, it should be considered that the linkage is affected by two types of errors: false links and missed matches. If the linkage errors are not properly taken into account, i.e. standard statistical procedures are applied to the linked data, biased estimates and mis-relationships between variables recorded in different sources may result. This paper provides a sensitivity analysis of the effect of linkage errors on the estimation of linear and logistic regressions. Different linkage scenarios are proposed, with various matching variables and accordingly different linkage error levels. The analysis confirms the importance of linkage errors and highlights the relevance of missed matches. The effectiveness of the proposed adjustment methods is demonstrated even when the conditions for their applicability are not fully satisfied, however a framework for taking into account the complexity of linkage procedures is needed.

Keywords: Unbiased estimators, probabilistic record linkage, regression, data integration, linkage errors

1. Introduction

The considerable effort to link data coming from different sources is not the objective of the statistical process but only an intermediate step. When dealing with data resulting from a record linkage process, it should be considered that the linkage can be affected by two types of errors: false links and missed matches.

In fact, these errors may affect the standard statistical analyses and if they are not properly taken into account, i.e. the standard statistical procedures are applied to the linked data, biased estimates and mis-relationships between variables recorded on different sources may result.

In recent years, increasing attention has been paid to tailored estimation procedures to take into account linkage errors. This paper aims at answering the question about the conditions for the negligibility of linkage error effects on estimation. The effects of different error levels are analysed in a simulated setting that re-

produces real linkage procedures. We propose a sensitivity analysis of the impact of linkage errors on linear and logistic regressions, we assume different linkage scenarios, with various matching variables characterised by different degrees of identifying power.

The paper is organized as follows: Section 2 reviews the effect of linkage errors on the total survey errors and briefly describes the probabilistic record linkage. Section 3 provides a brief account of the literature on statistical inference in the presence of linkage errors. In Section 4 the results of the sensitivity analysis are reported and discussed; finally, in Section 5, the outstanding issues are left open for future research.

2. Linkage errors and total survey errors

In a context where the integration of sources has acquired a preeminent role, the relevance of considering linkage errors in the total survey error representation has been acknowledged. See, for example, the extensions of the life cycle model of a statistical survey in Groves et al. [1] proposed by Bakker [2] and Zhang [3] for the integration of administrative data. Zhang [3] il-

*Corresponding author: Tiziana Tuoto, Italian National Statistical Institute, Istat Via Cesare Balbo 16, 00184 Rome, Italy. Tel.: +39 06 4673 3351; E-mail: tuoto@istat.it.

lustrates a representation where the linkage process is recognised as a possible cause of identification errors.

The total survey error in Biemer [4] classifies the various causes of errors. In his schema, the linkage procedures, as a step in data processing (see GSBPM v. 5), affect both measurement and frame errors.

In this paper, the attention is focused on the effect of linkage errors on relationships between variables, recorded in different sources. The probabilistic linkage process that generates linkage errors is briefly outlined in the next subsection.

2.1. *The probabilistic record linkage*

The fundamental theory of probabilistic record linkage is given by Fellegi and Sunter [5]. Given two lists, say L1 and L2, of size N_1 and N_2, let $\Omega = \{(a, b), a \in \text{L1 and } b \in \text{L2}\}$ be the set of all possible pairs, whose size is $|\Omega| = N_1 \times N_2$. The process of linking lists L1 and L2 can be seen as a classification problem where the pairs in Ω have to be assigned to two independent and mutually exclusive subsets M and U, such that:

M is the link set ($a = b$)

U is the non-link set ($a \neq b$).

In order to assign pairs to the sets M or U, a number, say K, of common identifiers, called the linking variables, are selected and, for each pair of records, a comparison vector $\gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_K\}$ is obtained. The ratio r between the conditional probability of γ given that the pair belongs to the set M and the conditional probability of γ given that the pair belongs to the set U

$$r = \frac{P(\gamma|(a, b) \in M)}{P(\gamma|(a, b) \in U)} = \frac{m(\gamma)}{u(\gamma)}$$

is used for classifying the pairs. The two probabilities m and u can be estimated, for instance as proposed by Jaro [6], assuming the true link status is a latent variable and using the EM algorithm. The pairs for which r is greater than an upper threshold T_m are assigned to the set of linked pairs, M^*; the pairs for which r is smaller than a lower threshold T_u are assigned to the set of unlinked pairs U^*; when the ratio r falls within the range (T_u, T_m), no decision is made and the pair is resolved by a clerical revision.

The thresholds T_u, T_m are chosen to minimize the false link probability, η, and the false non-link probability, $1 - \alpha$, defined as follows:

$$\eta = \sum_{\gamma \in \Gamma} u(\gamma) P(M^*|\gamma) = \sum_{\gamma \in \Gamma_{M^*}} u(\gamma)$$

$$\text{where } \Gamma_{M^*} = \{\gamma \colon T_m \leqslant m(\gamma)/u(\gamma)\} \quad (1)$$

$$1 - \alpha = \sum_{\gamma \in \Gamma} m(\gamma) P(U^*|\gamma) = \sum_{\gamma \in \Gamma_{U^*}} m(\gamma)$$

$$\text{where } \Gamma_{U^*} = \{\gamma \colon T_u \geqslant m(\gamma)/u(\gamma)\}. \quad (2)$$

The described linkage model also allows for the evaluation of the probability that a link is correct given that the link is assigned, the so-called true match rate:

$$\lambda = 1 - \frac{\sum\limits_{\gamma \in \Gamma_M} u(\gamma) P(M^*|\gamma)}{\sum\limits_{\gamma \in \Gamma_M} m(\gamma) P(M^*|\gamma)}$$

$$= 1 - \frac{\sum\limits_{\gamma \in \Gamma_{M^*}} u(\gamma)}{\sum\limits_{\gamma \in \Gamma_{M^*}} m(\gamma)}. \quad (3)$$

The parameter λ in Eq. (3) plays a fundamental role in the adjustment for linkage errors, as shown in the following section. However, the role of the probability of missing a link α is very relevant, as the results show in Section 4.3.

3. Methodologies for analyses on linked data

The effect of linkage errors on the linear regression model estimation was firstly illustrated by Neter et al. [7]. They show that even small linkage errors may produce a large bias in estimation procedures that do not tackle them. The original proposal for dealing with linkage errors [7] is subject to some restrictive assumptions: the two files have the same size and each record from one file is linked to a record in the second file with a constant probability λ of being correctly linked. Moreover, each record has the same probability $1 - \lambda$ of being incorrectly linked to any remaining record of the second file. Neter et al. [7] also generalise their basic model by assuming that the set of pairs from the two files can be partitioned into groups that meet the assumptions of constant probabilities.

Following their seminal paper, in recent years, many proposals have been suggested to provide sound inference in the presence of linkage errors. Scheuren and Winkler [8,9] and Lahiri and Larsen [10] extend the original work of Neter et al. [7] and propose methods for the unbiased estimation of linear regression coefficients under probabilistic record linkage, applying a bias correction to the Ordinary Least Squares (OLS) estimates.

In particular, Scheuren and Winkler [8,9] propose a ratio-type correction of the bias of the standard estimator on the basis of two critical assumptions: the proba-

bility of being a true match is known for each pair; the true match is the pair with the highest matching weight (probability).

Lahiri and Larsen [10] estimate the regression model between the linked values and the auxiliary variables. Their estimator still depends on the assumption of homoscedasticity. However, this condition generally is not satisfied, therefore, to relax this assumption, Chambers [11] suggests a Best Unbiased Estimator (BLUE) or its empirical (EBLUE) version.

Extensions to generalized linear models by means of generalised estimating equations (GEE) are proposed in Chambers [11] and Chambers et al. [12]. The GEE method [11] is subject to the same strong conditions as in the linear case: both registers have to be complete and no duplicates occur. An exchangeable linkage errors model is assumed, at least into groups of records.

Besides the EBLUE, Chambers [11] proposes a maximum likelihood (ML) estimator with application to the linear and logistic regression.

Finally, Chambers [11] considers the cases where the linkage is incomplete, i.e. one of the registers is a subset of the other register, as it is common in real data applications. The previous estimators can be extended to these cases under the assumption of no interaction between the sample selection process and the linkage error one, via weighted estimating functions. Kim and Chambers [13] apply the estimating equations as in Chambers [11] to deal with unlinked data and non-ignorable linkage models. Further analyses are needed in this context due to the strong assumptions and the limitation to the linear regression.

In the same setting of Chambers [11], Samart [14] extends the method to the class of linear mixed models. These models are very useful in the context of dependent observations, to take into account of intra-correlation of clustered units, e.g. students in a school, patients in a hospital, and they are largely exploited in Official Statistics for small area estimation. Proposals for small area estimation with linked data are in [15,16].

Finally, Chipperfield et al. [17] develop a ML approach for the analysis of probabilistically-linked records. The estimation technique is simple and it is implemented using the well-known EM algorithm. This method removes the limitation that all records have to be linked. This is a very important extension for dealing with administrative data when the different sources do not contain the same units or a file is not a subset of the other, as assumed in [11]. Moreover, their method explicitly considers both unlinked data

and missed links. Furthermore, unlike in [7,11] and the extensions discussed above, the method does not require exchangeability of linkage errors, even in groups of records. Therefore, it can also be applied when the linkage runs in several steps, as it is very frequent in real applications. They illustrate the method both for the analysis of contingency table and the logistic regression.

In the Bayesian approach, Fortini et al. [18] propose a different perspective to probabilistic record linkage. The objective of the inference is a linkage matrix C of size $N_1 \times N_2$, with $c_{i,j} = 1$ for cells corresponding to linked records and 0 otherwise. This formulation straightforwardly enables the inclusion of constraints on multiple matches, that are common in real applications and that can only be addressed with a separate optimization step in the Fellegi-Sunter approach [6]. An accurate description of the Bayesian model for record linkage is in [18,19].

In the Bayesian approach, the inference can be carried out at the same time of the record linkage procedure, i.e. the relationships between variables are estimated via the MCMC process simultaneously with the linkage model.

This means that at each iteration $t = 1, \ldots, T$ of the MCMC algorithm, a statistical analysis (e.g. a linear regression) is obtained on the basis of the match produced in $C(t)$.

The process causes a feed-back propagation of the information between the record linkage parameters and the more specific target quantities; i.e. the regression model depends on the selected matches, but even the selection of potential links depends on the information carried by the regression model. Tancredi and Liseo [19] illustrate the idea of feed-back propagation only for multiple linear regression, however no limitation prevents the application of the method to more general models. Nevertheless, as far as our knowledge, this method is computationally very costly and it is hardly practical for high-dimensional problems.

More recently, Steorts et al. [20] propose an alternative Bayesian approach that allows linking records from multiple lists simultaneously and at the same time de-duplicating the lists. The linkage is formulated as the process of recognising latent "entities" with a graphical representation, i.e. each record in the lists can be linked to a latent unit from 1 to $N_{\max}$, where $N_{\max}$ is the total number of units in all the lists, if no unit is present more than once.

4. A sensitivity analysis

The previous section summaries the increasing attention to linkage errors in statistical analyses on probabilistically linked data. However, the analyst may wonder whether to adopt sophisticated estimation procedures to adjust for linkage errors, or whether there are levels of linkage errors that can be ignored in subsequent analyses. Winkler [21] notes "Scheuren and Winkler [9] observed that, if linkage error is below 1%, then can perform statistical analysis without adjustment. Most 'good' matching situations have overall linkage error above 10%. Even 'high match scores' sets of pairs may have linkage error in range 1–5%. The current models may adjust the 'observed' matched pairs to having linkage error down from 10% to 7.5%. Bringing in sophisticated models that include edit/imputation may lower observed error to 5%. Further improved models may drop observed linkage error to 2.5%."

This paper aims at investigating whether linkage error adjustment is necessary, showing the effects of different error levels produced by real linkage procedures. To this purpose, we conduct a sensitivity study to analyze the impact of linkage errors in different scenarios in terms of match rates and linkage errors. We analyse the effectiveness of the most common adjustment methods for linear and logistic regressions, referred to in Section 3, with the aim of highlighting the gain in accuracy due to the adjustment of linkage errors. The applied estimators are described in Subsection 4.1. The fictitious population used for the sensitivity analysis is introduced in Section 4.2.

4.1. Adjusted estimators for linkage errors in regression models

Let y be the target variable and X be the matrix of auxiliary variables observed on the same units, assuming a linear relationship between the target variable and the covariates $y = X\beta + \varepsilon$, where β is the vector of regression coefficients and ε is the vector of i.i.d. random variables with $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$.

Due to linkage, the observed target variable is y^* instead of true unobserved y. Then, the standard estimator of the regression coefficient applied to the linked data is:

$$\hat{\beta}_{Naive} = (X^T X)^{-1}(X^T y^*). \tag{4}$$

Chambers [11] defines an exchangeable linkage errors model by assuming that the probability of correct linkage is the same for all records, or at least the probability is the same in groups partitioning the whole observations.

Chambers [11] models the relationship between the probabilistically linked data and the true data: $y^* = Ay$, where A is an unknown random permutation matrix, under the following assumptions:

1. The exchangeability of linkage errors;
2. The linkage is complete, i.e. the X and y refer to the same population and have no duplicates, so the two lists have the same size or the smallest list (e.g. a sample) reporting y, is contained in the largest one reporting X;
3. The linkage is one-to-one between the two lists.

As mentioned in Section 3.1, Scheuren and Winkler [8] propose a biased corrected OLS estimator of β, formulated by Chambers [11] as:

$$\hat{\beta}_{SW} = (X^T E X)^{-1}(X^T y^*) \tag{5}$$

where the matrix E is the expected value of the matrix A.

Alternatively, modeling the relationship between the linked data and the covariates, under the assumption of homoscedasticity, Lahiri and Larsen [10] propose the following estimator:

$$\hat{\beta}_{LL} = (X^T E^T E X)^{-1}(X^T E^T y^*). \tag{6}$$

To take also into account the heterogeneity, Chambers [11] proposes a Best Linear Unbiased Estimator (BLUE) of β:

$$\hat{\beta}_C = (X^T E^T \Sigma^{-1} E X)^{-1}(X^T E^T \Sigma^{-1} y^*) \tag{7}$$

where $\Sigma = \sigma^2 I + V$ is the variance matrix of y^*, and V is the variance $Var(AX\beta)$.

For the linear logistic model, $\text{logit}(p) = X\theta$, where p is the expected value of a binary variable y, a GEE approach is applied in Chambers [11]; the bias corrected form of the estimating functions is

$$H_{adj} = G(\theta)(y^* - Ef(\theta)), \tag{8}$$

where $f(\theta) = E_X(y)$ and $G(\theta)$ is a function of θ and X.

Under perfect linkage, the logistic model is fitted via ML with $G(\theta) = X^T$ then Eq. (8) can be specified as:

$$H_M = X^T(y^* - Ef(\theta)). \tag{9}$$

Alternatively, the function $G(\theta) = X^T E^T$ corresponding in the linear case to the Lahiri and Larsen estimator in Eq. (6), leads to:

$$H_{LL} = X^T E^T(y^* - Ef(\theta)), \tag{10}$$

and $G(\theta) = X^T D(\theta) E^T \Sigma^{-1}(\theta)$, corresponding in the linear case to the BLUE of Eq. (7), leads to:

$$H_C = X^T D(\theta) E^T \Sigma^{-1}(\theta)(y^* - Ef(\theta)). \quad (11)$$

In the following, we denote the estimators from Eqs (9)–(11) as M, LL and C respectively.

4.2. Experimental data

For the sensitivity analysis, the fictitious population generated by the data from the ESSnet DI [22] is used; the ESSnet DI was a European project on data integration (Record Linkage, Statistical Matching, Micro integration Processing) run from 2009 to 2011. The data are freely available online at http://web.archive.org/web/20150930225743/http://www.cros-portal.eu/sites/default/files//Transfer%20to%20Istat.zip (accessed 31 January 2018). The ESSnet DI provides a number of fictitious data sources, which are supposed to have captured details of persons at the same reference time. In these data sets, which comprise over 26000 records each, linking variables (names, dates of birth, addresses) for individual identification may be distorted by missing values and typos, to imitate real-life situations. These synthetic data reproduce the real data and the actual observed errors that make the linkage procedure complex. This is a key point of this study, where for the first time regression model adjustments for linkage errors are analysed under realistic linkage procedures. In this simulation set-up the true match status is known and hence the linkage results can be benchmarked to evaluate the true linkage errors.

The ESSnet DI datasets are augmented with an explanatory and dependent variables. For the linear regression model, the variables are generated according to the following model as in Chambers [11]:

$$X \sim [1, \text{Uniform}(0, 1)]$$
$$Y = X\beta + \varepsilon$$
$$\varepsilon \sim N(0, 1)$$
$$\beta = [1, 5]$$

For the logistic model, the variable X is independently generated for each record assuming values $\{1, 2\}$ with probabilities $p_X = \{0.75, 0.25\}$. The variable $Y = \{1, 2\}$ is generated from the multinomial distribution with parameter (0.7, 0.05, 0.2, 0.05), i.e. $p(Y|X = 1) = \{0.93, 0.07\}$ and $p(Y|X = 2) = \{0.8, 0.2\}$.

For this analysis, 100 samples of size 1000 are generated, sampling the data independently and randomly

Table 1
Results of linkage procedures in the three scenarios

Scenario	Average true matches	Average declared links	Average false matches rate (%) $1 - \lambda$	Average missing matches rate (%) $1 - \alpha$
Gold	858	807	0.006	6.031
Silver	858	730	1.868	14.91
Bronze	858	744	2.016	16.70

without replacement. Then from each sample, two different lists are generated mimicking the undercoverage of register data and the presence of errors in identifiers. The coverage rates of the two lists, say L1 and L2, are 0.93 and 0.92 respectively. The auxiliary variable and the target variable are separately assigned in list L1 and L2 respectively.

Three different linkage scenarios are proposed. In the first scenario, the Gold scenario, the variables with the highest identifying power are used as linking variables (Name, Surname, Complete date of birth). A second scenario, the Silver scenario, is considered applying the linkage with less variables (Complete date of birth, i.e., Day, Month and Year of Birth). Finally, the third scenario, the Bronze scenario, uses variables with the least identifying power, being more affected by errors than the others are (Surname, Day and Month of Birth).

The probabilistic record linkage is performed, according to the Fellegi and Sunter theory [5] as implemented in the RELAIS software [22].

Table 1 summaries the results of the linkage procedures, averaging on the 100 replicas, the table reports the number of the true matches, the number of declared links by the procedure, the percentage of false match rate $(1 - \lambda)$, the percentage of missing matches rate $(1 - \alpha)$.

The Gold scenario gives almost no false matches but only missing matches. On the other hand, the Silver and Bronze scenarios result in both false matches and missing matches, with different error rates.

As the true matching status is known, the true error rates $(1 - \lambda)$ and $(1 - \alpha)$ can be evaluated comparing the true matches and the realised links:

	True matches	True un-matches
Links	a – true positives	b – false positives or false links
No-links	c – false negatives or missing links	d – true negatives

The quantity $(1 - \lambda)$ is defined as the ratio between the false links declared by the linkage procedure and the overall number of declared links and the quantity

Table 2
Linear model – naive and adjusted estimators

Linkage scenario	Estimator	Intercept	Standard error for intercept	Slope	Standard error for intercept
Population	True value	1.01	0.0122	4.97	0.0212
Samples	Under perfect linkage	1.009	0.0655	4.978	0.1140
Gold	Naïve	1.021	0.0705	4.955	0.1223
	SW-LL/C	1.021	0.0704	4.955	0.1221
Silver	Naïve	1.058	0.0761	4.868	0.1321
	SW-LL	1.012	0.0779	4.961	0.1363
	C	1.013	0.0778	4.960	0.1362
Bronze	Naïve	1.073	0.0769	4.848	0.1335
	SW-LL	1.023	0.0789	4.948	0.1381
	C	1.023	0.0788	4.947	0.1380

Average values on 100 replicates.
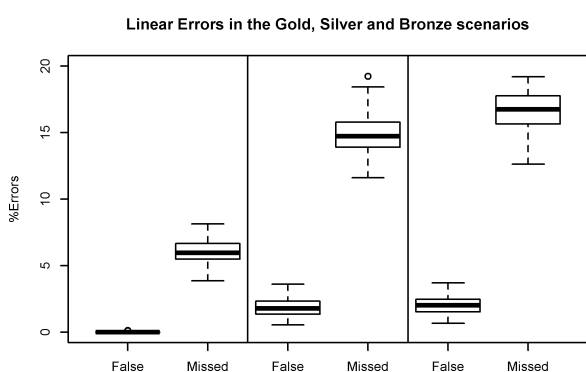
Fig. 1. Linkage errors in the three Scenarios.

Table 3
Logistic model – naive and adjusted estimators

Linkage scenario	Estimator	Beta	Standard error	Coverage
Population	True value	−1.57	0.016	
Perfect linkage	Naïve	−1.584	0.0863	
Gold linkage	Naïve	−1.577	0.0924	85
Silver linkage	Naïve	−1.579	0.0963	84
	M	−1.582	0.0967	85
	LL	−1.588	0.0971	85
	C	−1.599	0.0976	84
Bronze linkage	Naïve	−1.575	0.0971	84
	M	−1.578	0.0976	85
	LL	−1.585	0.0979	86
	C	−1.596	0.0984	83

Average values on 100 replicates.

$(1 - \alpha)$ is defined as the ratio between true matches missed by the linkage procedure and the overall number of true matches.

The error $(1 - \lambda)$ plays an important role in the adjustments. The use of the true values for $(1 - \lambda)$ and $(1 - \alpha)$ instead of estimating them (e.g. via Eqs (2) and (3)) allows the comparison of the adjusted regression estimators against the naïve estimator without the effect of the linkage errors estimation. The distribution of the False Match Rate and the Missing Match Rate over the 100 replicates is represented in Fig. 1.

The three scenarios produce increasing values of false and missing match rates. Primarily, the false match rate is under control, as in real practical applications where a conservative linkage strategy is often preferred; indeed, in our worst scenario the values of the false match rate are less than 5%. On the other hand, the control of the false match rate results in an increase of the missing match rate, which takes up to 20% in the Bronze Scenario.

However, under the standard assumption of ignorability of the missing linkage mechanism, one expects that the presence of missing matches does not introduce bias in the estimates even if it has an impact on their variability.

4.3. Results

The following tables show the results of the previous estimators Eqs (4)–(7) for the linear model as well as of the standard estimator and the derived estimators from Eqs (9)–(11) for the logistic model.

As already mentioned in Section 3, the assumptions of the methods proposed in Chambers [11] are far to be met in practical situations. In our simulation, although different real contexts are reproduced, the exchangeability is assumed for the application of the adjusted estimators. Even if the assumption of exchangeability of linkage errors is not met even in sub-groups, only one block is considered and an overall value of the probability of correct match λ is applied for correcting the bias.

Table 2 reports the values of the linear regression model parameters and their relative standard errors for the naïve (4) and the adjusted estimators, SW (5), LL (6) and C (7), as well as the true values calculated on the whole population of 26450 records and the estimates obtained with the sample under perfect linkage.

In Fig. 2, the percentage relative errors of the estimates under perfect linkage (TR), the naïve estimates
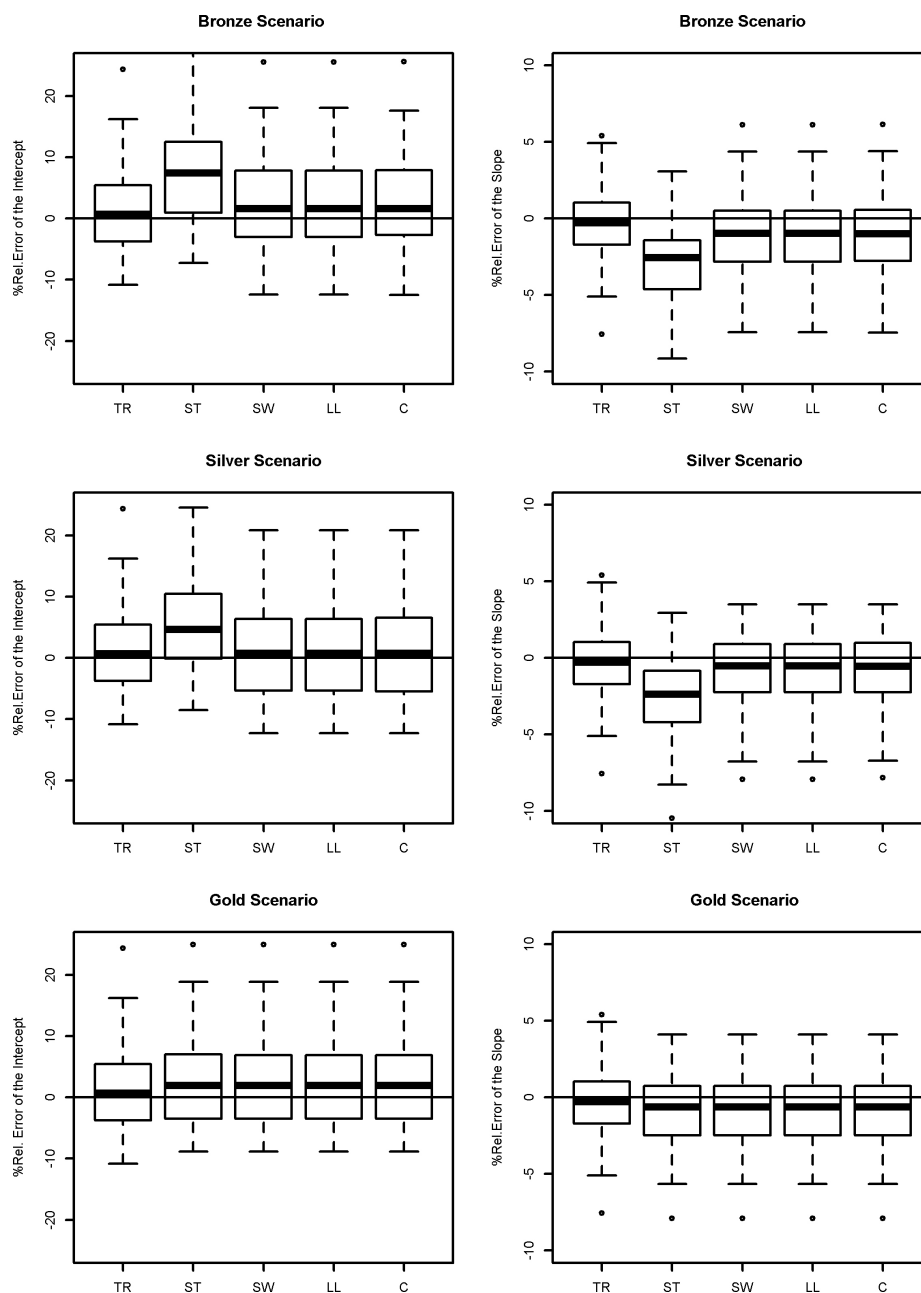
Fig. 2. Percentage of relative errors for true values, standard estimators and adjusted estimators in linear regression model.

in case of linkage errors (N) and the adjusted estimates in linear regression model are plotted.

As expected, the presence of false matches weakens the correlation between x and y in the linear regression model. Hence, the naïve estimator of the slope is increasingly biased towards 0, as the false linkage rate increases. On the contrary, the estimator of the intercept is biased in the opposite direction (if the mean value

of the observed y, y^* does not change). Even if the exchangeability assumption is not valid, the adjusted estimators provide bias reduction when the false match rate is not negligible with a limited increase in the variance compared to the naïve estimator.

The results of the logistic regression model are summarised in Table 3, where the parameter estimates and their relative standard errors are reported, averaging on
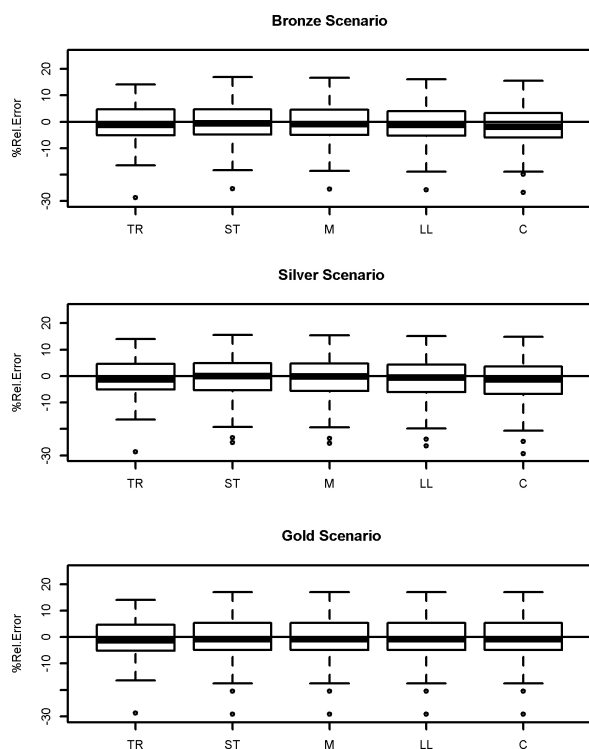
Fig. 3. Percentage of relative errors for true values, standard estimators and adjusted estimators in logistic regression model.

100 replicas. The table also shows the number of time the confidence intervals of the obtained estimates contains the true parameter value, i.e. the coverage of the nominal 95% CIs.

In Fig. 3 the distributions of relative errors of parameter estimates are plotted.

For categorical data, the false match error has less impact because of the nature of the response variable: the values of y and y^* can be the same even when there is a linkage error and the conditional distribution $(y^*|x)$ may not differ appreciably from the conditional distribution $(y|x)$. The results show that the bias of the naïve estimator of β does not increase noticeably from the Gold to the Bronze scenario. In this case, the correction does not reduce the bias of the standard estimator.

The results of the coverage evaluation show that the standard error evaluation might be severely affected by underestimation.

These results suggest that the missing matches should also be taken into account to completely remove the bias. Indeed, in the Gold scenario, where the false matches are close to zero, the naïve and the adjusted estimators are still biased due to a not-ignorable missing mechanism. The bias effect of the missing matches

is also shown in the other scenarios, where the adjustments for false matches reduce the bias but do not eliminate it. In any case, the correction for the bias is more effective in the linear than in the logistic model: in linear regression, we achieve a bias reduction of about 10% for the Bronze scenario and a little smaller in the Silver one. However, more work is needed for the estimation of the logistic regression, where the naïve estimator is slightly closer to the benchmark value than the adjusted ones, but about 15% of the replicas produces values out of the nominal 95% CI.

As expected, the comparison of the standard errors of the estimator under perfect linkage and the naïve estimator shows that the occurrence of missed matches also produces an increase in variance due to the reduction of the observed sample size, similarly to a missing value mechanism.

5. Concluding remarks

This work proposes a sensitivity analysis of the effect of linkage errors both on bias and variability of regression estimates, when linkage errors are assumed to be known. The need for the adjustment is evident even with small level of false matches. In our simulation, an average false match error of around 2% results in a percentage relative bias of the intercept and the slope of regression models respectively equal to 4.86% and -2.2% in the Silver scenario and 6.34% and -2.6% in the Bronze scenario.

In the linear case, the adjustment is effective in reducing the bias without an appreciable increase in the standard errors of the estimates, when the linkage errors are known. However, further analysis should be conducted to assess the trade-off between the adjustment of the bias and the expected increase in variance when one needs to estimate the linkage errors, as it is usual in practice. Actually, the linkage errors evaluation is not a straightforward task, some proposals are in Belin and Rubin [24] and Tuoto [25].

As observed above, the bias associated to missed matches is also substantial, since in practice the ignorability assumption of the linkage mechanism may be unmet; hence the role of the missing matches should be further analysed.

As shown in this work, the model proposed in Chambers [11] enables the reduction of the bias but it does not consider all the complexities of a real linkage procedure (see Sections 3 and 4). Chipperfield et al. [17] propose a model which is subject to less strin-

gent assumptions on linkage errors than the Chambers's model. Their proposal does not require the exchangeability of linkage errors and considers explicitly the erroneous missed matches in the adjustment. It would be interesting to compare the two approaches in a simulation setting based on a real linkage procedure.

Acknowledgments

We wish to thank the Editor-in-Chief Kirsten West for encouraging us to improve an earlier version of the manuscript and Prof. Li-Chun Zhang for useful comments and discussion.

References

[1] Groves RM, Fowler FJ, Couper MP, Lepkowski JM, Singer E, Tourangeau R. Survey Methodology, 2nd Edition, Wiley 2004.

[2] Bakker B. Micro-integration: State of the Art. Paper for the 2010 Joint UNECE/Eurostat Expert Group Meeting on Register-Based Censuses, The Hague, The Netherlands.

[3] Zhang L-C. Topics of statistical theory for register-based statistics and data integration. Statistica Neerlandica 2012; 66.

[4] Biemer P. Total survey error design, implementation, and evaluation. Public Opinion Quarterly 2010; 74(5): 817-848.

[5] Fellegi IP, Sunter AB. A Theory for record linkage. Journal of the American Statistical Association 1969; 64: 1183-1210.

[6] Jaro M. Advances in record linkage methodology as applied to matching the 1985 test census of Tampa, Florida. Journal of American Statistical Association 1989; 84: 414-420.

[7] Neter J, Maynes ES, Ramanathan R. The effect of mismatching on the measurement of response error. Journal of the American Statistical Association 1965; 60: 1005-1027.

[8] Scheuren F, Winkler WE. Regression analysis of data files that are computer matched – part I. Survey Methodology 1993; 19: 39-58.

[9] Scheuren F, Winkler WE. Regression analysis of data files that are computer matched – part II. Survey Methodology 1997; 23: 157-165.

[10] Lahiri P, Larsen M. Regression analysis with linked data. Journal of the American Statistical Association 2005; 100: 222-230.

[11] Chambers R. Regression Analysis Of Probability-Linked Data. Official Statistics Research Series 2009; 4.

[12] Chambers R, Chipperfield J, Davis W, Kovacevic M. Inference Based on Estimating Equations and Probability-Linked Data. Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 18-09, 2009. http://ro.uow.edu.au/cssmwp/38.

[13] Kim G, Chambers R. Regression analysis under incomplete linkage. Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 17-2009.

[14] Samart K. Analysis of probabilistically linked data. Doctor of Philosophy Thesis, School of Mathematics and Applied Statistics, University of Wollongong, 2011. http://ro.uow.edu.au/theses/3513.

[15] Di Consiglio L, Tuoto T. Small Area Estimation in the Presence of Linkage Errors. In: Ferraro M. et al. (eds) Soft Methods for Data Science, Advances in Intelligent Systems and Computing Springer, Cham 2017; 456.

[16] Briscolini D, Di Consiglio L, Liseo B, Tancredi A, Tuoto T. New methods for small area estimation with linkage uncertainty. International Journal of Approximate Reasoning 2018; 94: 30-42.

[17] Chipperfield JO, Bishop GR, Campbell P. Maximum likelihood estimation for contingency tables and logistic regression with incorrectly linked data. Survey Methodology 2011; 37(1): 13-24.

[18] Fortini M, Liseo B, Nuccitelli A, Scanu M. On Bayesian Record Linkage. Research in Official Statistics 2001; 1: 185-198.

[19] Tancredi A, Liseo B. Some advances on Bayesian record linkage and inference for linked data, workshop Essnet DI, Madrid, November 2011.

[20] Steorts R, Hall R, Fienberg SE. A Bayesian approach to graphical record linkage and de-duplication. Journal of the American Statistical Association 2015. URL: http://arxiv.org/abs/1312.4645.

[21] Winkler WE. Quality and analysis of national files – computational methods for censuses and surveys. Presentation, January 9, 2014.

[22] Essnet DI, McLeod P, Heasman D, Forbes I. Simulated data for the on the job training, 2011. https://ec.europa.eu/eurostat/cros/content/job-training_en.

[23] RELAIS. User's guide version 3.0, 2015, available at https://joinup.ec.europa.eu/release/relais-30-0.

[24] Belin TR, Rubin DB. A method for calibrating false-match rates in record linkage. Journal of the American Statistical Association 1995; 90: 694-707.

[25] Tuoto T. New proposal for linkage error estimation. Statistical Journal of the IAOS 2015; 32(2).