When adjusting for bias due to linkage errors: a sensitivity analysis

Loredana Di Consiglio, Tiziana Tuoto

Istat – Istituto Nazionale di Statistica

Via Cesare Balbo 16, 00184 Roma

e-mail: diconsig@istat.it , tuoto@istat.it

**Abstract:**

The huge accomplished effort to link data is not the final aim of the statistical process. However, whatever is the statistical analysis to be performed on such an integrated data, when dealing with data resulting from a record linkage process, it should be taken into account that linkage is subject to two types of errors: erroneous acceptance of false links and rejection of true matches. In fact, these errors may affect the standard statistical analyses and if they are not properly accounted for, i.e. naïve procedure are applied to linked data, biased estimates and misreported relationship among variables from different sources may result. In this work a sensitivity analysis on the effect of linkage errors on estimation of linear and logistic regressions has been carried out assuming different linkage scenarios with matching variables of different degree of identifying power. Relevance of erroneous un-matches is highlighted as well as need for framework for taking into account complexities of the linkage procedures.

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

**1. Introduction**

The huge accomplished effort to link data is not the final aim of the statistical process. However, whatever is the statistical analysis to be performed on such an integrated data, when dealing with data resulting from a record linkage process, it should be taken into account that linkage is subject to two types of errors: erroneous acceptance of false links and rejection of true matches.

In fact, these errors may affect the standard statistical analyses and if they are not properly accounted for, i.e. naïve procedure are applied to linked data, biased estimates and misreported relationship among variables from different sources may result.

In this work a sensitivity analysis on the effect of linkage errors on estimation of linear and logistic regressions has been carried out assuming different linkage scenarios with matching variables of different degree of identifying power. Section 2 introduces the effect of linkage errors on the total survey errors. Section 3 gives a brief introduction of the literature on the issues of producing sound statistical inference in presence of linkage errors. In section 4 the results of the sensitivity analysis is reported and discussed and finally in section 5 further work is envisaged.

**2. Linkage Errors and Total Survey Errors**

In a context where integration of different sources has acquired a preeminent role, the relevance of considering linkage errors in the total survey error representation has now recognized. See for example Bakker (2010) and Zhang (2012) that propose an extension of the Groves et al. (2004) life cycle model of a statistical survey. The following Figure illustrates the representation proposed in Zhang 2012 where the process of linkages of different sources may produce identification errors.



The following representation of total survey error in Biemer (2010) classify different causes of errors, in this case linkage procedures – a process that may be, as in GSBPM v. 5 is considered as included in the data processing, have a side effect both in measurement errors and in the frame errors.

 

In this work focus is given to the effect of linkage errors in measurement, being multiple variable a by-product of the integration of different sources containing separatly its components.

The effect of linkage errors on the evaluation of parameters in linear regression models for variables reported in different sources was firstly illustrated by Neter, Maynes and Ramanathan (1965). They show how even small errors may greatly affect the results of the estimation procedures that do not tackle the linkage errors. Following their seminal paper, in recent years, many proposals have been suggested to provide sound inference in presence of linkage errors. The following section gathers a short account of literature.

**3. Methodologies for analyses on linked data**

The original proposal of Neter, Maynes and Ramanathan (1965) is subject to some restrictions: it is assumed that the two files have the same size and each record from one file always could be matched to a record in the other file with the same probability p of being correctly matched. Moreover, each record has the same probability q of being mismatched to any remaining records of the other file. They also generalized their basic results by assuming that set of pairs from the two files could be partitioned into groups in which the previous assumptions hold.

Scheuren and Winkler (1993, 1997) and Lahiri and Larsen (2005) extend the original work of Neter et al. (1965) and proposed methods to calculate unbiased estimates of coefficients for a linear regression model under probabilistic record linkage, applying a bias correction to the Ordinary Least Squares (OLS) estimates. In particular, the proposal from Scheuren and Winkler (1993, 1997) is based on a ratio-type correction for the bias in the standard estimator and the adjustment is based on two crucial assumptions, first the probability of being a true match is available for each pair, second, the true match is the pair with the highest matching weight (probability). On the other hand, the optimality of the proposal from Lahiri and Larsen (2005) depends on homoscedasticity of regression errors of the linear model with modified set of explanatory variables that take into account linkage errors. This condition generally does not hold, therefore Chambers (2009) suggests a Best Unbiased Estimator (BLUE) or its empirical (EBLUE) version.

Extensions of possible applications to generalized linear models by applying generalised estimating equations are proposed in Chambers (2009) and Chambers, Chipperfield, Davis and Kovacevic (2009. The method in Chambers (2009) is subject to the same strong conditions as linear case: both registers are complete and no duplicates occur, i.e. each register contains all the N records of the population. An exchangeable linkage errors model is assumed, at least into groups of records. Besides EBLUE, Chambers (2009) proposes a maximum likelihood (ML) estimator under the assumption of normality with application to linear and logistic regression functions.

An important extension considered by Chambers (2009) consists in assuming that the two sources do not contain an equal number of records that link according to a 1:1 basis, but it allows dealing with the situation where linkage is incomplete in the sense that one of the registers in fact only covers a subset of the population fully covered by the other register, as commonly occurs in real data application. He proposes to estimate the model on the basis of the linked sample data only under the hypothesis of no interaction between sample selection process and linkage error processes, using weighted estimating function. Kim e Chambers (2009) consider also the unlinked data and non-ignorable linkage model in the framework of estimating equations, exploiting the signs of the regression errors for linked records. Further analyses are needed in this field due to the strong assumptions and the restriction to the linear regression of their study.

In the same setting, Samart (2011) extends the research to the class of linear mixed models with linked data. This class of models is very useful in the context of dependent observations, for example units that belongs to the same sanitary structure as in the following ISPO application, to take into account of intra-correlation effects.

Finally, Chipperfield, Bishop and Campbell (2011) develop a ML approach for analysis of probabilistically-linked records. The estimation technique is simple and is implemented using the well-known EM algorithm. This method overcomes the restraint that all records have to be linked and explicitly consider in the analysis the unlinked data or missed links. This constrain removal is particularly important when dealing with administrative data where the different sources do not contain exactly the same units or the set of units on one file in not a subset of those on the other file.

Furthermore, the method does not require exchangeability of linkage errors, even in groups of records (unlike in Neter et al., 1965 and in Chambers, 2009), then it can be applied without restrictions also to cases when multiple steps are performed for linking records, as it is very common in real-world applications. They illustrate the method both for the analysis of contingency table and logistic regression.

Finally, we recall the work of Liseo and Tancredi (2011). They study the analysis of probabilistically linked data in a fully Bayesian framework. In their approach all the uncertainty about the matching process is automatically accounted for in the subsequent inferential steps. Fortini et al. (2001) propose a Bayesian model to estimate the whole linkage matrix C, containing 1s on cell corresponding to linked records and 0s otherwise. This framework allows considering in the definition of the problem constraints on multiple matches, that are common in real applications and that are treated in a separate optimization step in the Fellegi-Sunter approach (Jaro, 1989).

A complete description of the Bayesian model for record linkage is in Fortini et al. (2001) and in Tancredi and Liseo (2011), a description of the MCMC method to obtain the estimate of the condition posterior distribution is given, as well.

In the Bayesian approach, the natural way to perform the inferential analysis is to carry that out at the same time of the record linkage procedure, i.e. including the analyses of relationships into the MCMC process for estimating the linkage model.

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

This will cause a feed-back propagation of the information between the record linkage parameters and the more specific target quantities. In their paper they propose this idea for estimating multiple linear regression, however no particular condition prevents to apply it in more general problems. As far as our knowledge, this method is computationally very costly and it is hardly practical for high-dimensional problems.

**4. A sensitivity analysis**

Winkler 2014 notes “Scheuren and Winkler (1997) 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%.” In the following we report a sensitivity study to analyze the impact of linkage errors with fictitious data reproducing different linkage scenarios in terms of levels of match rate and linkage errors.

*4.1. Experimental data*

A population of 1000 units have been randomly sampled from the fictitious population census data created by Paula McLeod, Dick Heasman and Ian Forbes, ONS, for the ESSnet DI (McLeod, Heasman and Forbes, 2011). Two sets of explanatory and dependent variables were generated in the population, replicating simulations as in Chambers (2009) for linear model:

X~[1,Uniform(0,1)] Y= Xβ+ε ε~Norm(0,1) β=[1,5]

As far as set of categorical variables, for the logistic model, the variable X was generated independently for each record such that x =1 with probability 0.75 and x =2 otherwise. The variable y can take the values 1 or 2 and was generated from the multinomial distribution with parameter (0.7,0.05,0.2,0.05).

Two different lists were generated mimicking the process of undercoverage of register data and the presence of errors in identifiers. Different subsets of variables of the population were assigned to the two different lists. The two lists, say L1 and L2, include 942 and 921 units respectively. Units in common (the true matches) are 868 whereas true un-matches are 127. In L1 explanatory variables were reported while L2 contained the dependent one.

Three different linkage scenarios were illustrated. Probabilistic record linkage procedures, according to the seminal theorization by Fellegi and Sunter (1969), were applied by means of the software RELAIS (2011). In the first scenario, Gold scenario, the variables with the highest identifying power were used for the linkage (Name, Surname, Complete date of birth). This allowed to declare 826 matches without including any false match.

A second scenario, Silver scenario, was proposed applying the linkage on less variables (Name, Surname, Year of Birth). This allowed to declare 752 matches including 11 false matches.

Finally the third scenario, Bronze scenario, used variables less identifying power (Day of birth, month of birth, address). This allowed to declare 786 matches including 30 false matches.

Table 1 summaries the result of the linkage procedures, illustrating the probability of missing true matches (λ), the probability of false match (μ) and the false match rate (η)

**Table 1 – Results of linkage procedures in the three scenarios**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Scenario | Declared Matches | False matches in Declared | λ | μ | η |
| Gold | 826 | 0 | 0.048 | 0 | 0 |
| Silver | 752 | 11 | 0.146 | 0.087 | 0.015 |
| Bronze | 786 | 30 | 0.129 | 0.236 | 0.038 |

where λ is the ratio between true matches missed by the linkage procedure and all the true matches, μ is the ratio between false matches declared by the linkage procedure and the total amount of non-matches and finally η is the ratio between the false matches declared by the linkage procedure and the overall number of declared matches.

*4.2 Linkage errors in regression models*

The following tables show the results of naive Estimators for linear and logistic models and those of estimators proposed in Chambers (2009).

As mentioned in section 3, the assumptions of the methods proposed in Chambers are far to hold in a practical situation. In this case, where a real context is simulated, some simplifying adjustments were necessary: as exchangeability of linkage errors does not hold even in sub-groups, only one block was considered and an overall value of the probability of correct match (1-η) was applied for correcting bias due to linkage errors. Moreover, bias adjustment could be evaluated only in Scenarios Silver and Bronze given that error related to missed links is not considered in Chambers (2009).

**Table 2 Linear Model – Naive Estimator linkage and Linkage error bias adjusted estimators**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Linkage scenario** | **Estimator** | **Beta** | **Standard** **Error** | **Lower Bound** | **Upper Bound** |
| **Population** | **True Value** | 0.886 5.155 | 0.064 0.112 | 0.760 4.936 | 1.012 5.373 |
| **Perfect Linkage** | **Naïve** | 0.907 5.093 | 0.069 0.121 | 0.772 4.856 | 1.043 5.329 |
| **Gold Linkage** | **Naïve** | 0.927 5.085 | 0.071 0.123 | 0.788 4.843 | 1.066 5.327 |
| **Silver Linkage** | **Naïve**  | 0.988 4.976 | 0.079 0.138 | 0.833 4.706 | 1.143 5.245 |
| **Ratio type****Modified OLS Predictive**  | 0.9525.050 | 0.080 0.141 | 0.794 4.774 | 1.109 5.326 |
| **Eb\_CUE**  | 0.9495.055 | 0.080 0.141 | 0.792 4.779 | 1.1075.331 |
| **Bronze Linkage** | **Naïve**  | 1.0454.876 | 0.078 0.135 | 0.892 4.611 | 1.197 5.142 |
| **Ratio type****Modified OLS Predictive** | 0.9495.070 | 0.081 0.144 | 0.789 4. 787 | 1.108 5.353 |
| **Eb\_CUE**  | 0.9475.075 | 0.081 0.144 | 0.7874.793 | 1.1065.358 |

**Table 3 Logistic Model - Naive Estimator and Linkage error bias adjusted estimators**

|  |  |  |  |
| --- | --- | --- | --- |
| **Linkage scenario** | **Estimator** | **Beta** | **Standard** **Error** |
| **Population** | **True Value** | -1.680 | 0.087 |
| **Perfect Linkage** | **Naïve** | -1.744 | 0.096 |
| **Gold Linkage** | **Naïve** | -1.762 | 0.100 |
| **Silver Linkage** | **Naïve**  | -1.795 | 0.106 |
| **M** | -1.798 | 0.106 |
| **A** | -1.803 | 0.107 |
| **C** | -1.817 | 0.107 |
| **Bronze Linkage** | **Naïve**  | -1.734  | 0.101  |
| **M** | -1.741  | 0.102 |
| **A** | -1.755  | 0.102 |
| **C** | -1.789  | 0.104 |

In table 3, adjusted estimators are recalled as in page 22 of Chambers (2009): shortly option M defines the estimating equation for the MLE under perfect linkage, option A leads to the Lahiri-Larsen estimator under a linear model and option C defines the second order efficient version of estimating function given in Chambers (2009) page 19.

The obtained results show the relevance of the missing matches and its error components should be taken into account to completely remove linkage errors effect on the estimate bias. In fact, the naïve estimators under perfect linkage and Gold scenario, where no false matches are included, are still biased due to missing true matches. This effect is also shown in the logistic regression (table 3), under the Bronze scenario the naïve estimate is closer than the adjusted ones to the true value because there the missed matches component is lower (see table 1) than in the other scenarios. The correction for bias is effective in the linear case (achieving a bias reduction of about 10% for the silver scenario and higher in the bronze one) but more work is needed for the logistic case where the Naïve estimator performs slightly better.

**5. Concluding Remarks**

In this work, a sensitivity analysis on the effect of the linkage errors on bias component of regression estimates are provided, further works are needed in order to investigate linkage errors effects on variability component. Further analyses should be done to assess the trade-off in adjusting for bias with respect to the expected increase of variance.

As observed in the last section the inclusion of the erroneous missed matches is substantial. Framework proposed in Chambers (2009) and here adopted does not consider to all the complexities (see sections 3 and 4) of a real linkage procedures; a framework less subject to strong assumptions is proposed in Chipperfield et al. (2011) where for instance exchangeability of linkage errors is not required and erroneous missed matches are explicitly considered in the adjustment. Finally, here the probability of being correctly linked and the probability of erroneous missed matches are assumed to be known, whereas the linkage errors evaluation is not a straightforward task (Belin and Rubin, 1995). A further sensitivity analysis will be conducted taking into account their proposal.

**References**

Bakker, B. (2010) *Micro-integration: State of the Art*. Paper for the Joint UNECE/Eurostat Expert Group

Meeting on Register-Based Censuses, The Hague, The Netherlands.

Belin, T. R., and Rubin, D. B. (1995). A Method for Calibrating False-Match Rates in Record Linkage. Journal of the American Statistical Association, 90, 694-707.

Biemer (2010) Total Survey Error Design, Implementation, And Evaluation Public Opinion Quarterly, Vol. 74, No. 5, 2010, pp. 817–848

Chambers, R. (2009). Regression Analysis Of Probability-Linked Data. Official Statistics Research Series 4.

Chambers, R., Chipperfield, J., Davis, W. and Kovacevic, M. (2009) 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

Chipperfield, J. O., Bishop, G. R . and Campbell P. (2011). Maximum likelihood estimation for contingency tables and logistic regression with incorrectly linked data Survey Methodology, June 2011 13 Vol. 37, No. 1, pp. 13-24

Fellegi I.P., Sunter A.B. (1969) “A Theory for record linkage”, Journal of the American Statistical Association, 64, 1183-1210.

Fortini, M., Liseo, B., Nuccitelli, A. and Scanu, M. (2001) On Bayesian Record Linkage. Research in Official Statistics, 4, Vol.1, 185-198.

Herzog, T. N., Scheuren, F. and Winkler, W. E. (2007) Data Quality and Record Linkage Techniques. Springer.

Kim, G. and Chambers, R., (2009) Regression analysis under incomplete linkage, Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 17-09.

Kim, G. and Chambers, R., (2012) Unbiased regression estimation for multi-linked data in the presence of correlated linkage error, Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 3-12.

Lahiri, P. and Larsen, M. (2005). Regression analysis with linked data. Journal of the American Statistical Association, 100, pp-222-230

McLeod, Heasman and Forbes, (2011) Simulated data for the on the job training, Essnet DI http://www.cros-portal.eu/content/job-training

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

RELAIS, (2011). User’s guide version 2.2, available at http://joinup.ec.europa.eu/software/relais/release/22

Samart, K. (2011). 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

Scheuren, F. and Winkler, W.E. (1993). Regression analysis of data files that are computer matched – Part I. Survey Methodology, 19, pp. 39-58.

Scheuren F. and Winkler W.E. (1997). Regression analysis of data files that are computer matched- part II, Survey Methodology, 23, pp. 157-165.

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

Winkler W.E. (1995), “Matching and Record Linkage”, in Cox, Binder, Chinnappa, Christianson, Colledge, Kott (eds), Businness Survey Methods, Wiley & Sons, pp. 355-384.

Winkler, W.E. (2014), Quality and Analysis of National Files - Computational Methods for Censuses and Surveys, Presentation, January 9, 2014

Winkler, W.E. (2011), Cleaning and using administrative lists: Enhanced practices and computational algorithms for record linkage and modeling/editing/imputation, Section on Survey Research Methods, JSM 2011

Zhang, L.-C. (2012), Topics of statistical theory for register-based statistics and data integration. Statistica Neerlandica, 66: