Contribution to the discussion of “Statistical challenges of administrative and transaction data”

Paul A. Smith, S3RI, University of Southampton, Highfield, Southampton, SO17 1BJ. p.a.smith@soton.ac.uk
Raymond L. Chambers, University of Wollongong, ray@uow.edu.au

We congratulate Professor Hand on a paper that poses challenges across a wide range of topics in administrative data; here we focus on analysis of linked data and implications for a general framework (challenges 12 and 13).

Analysis of linked data is now widespread, enabled by data sharing legislation such as in the Statistics and Registration Service and Digital Economy Acts, and also by projects like the Administrative Data Research Network. Some linkages use unique identifiers, but often linkage is probabilistic, based on matching of record-level characteristics, and hence is subject to error.

Given N potential matches and a correct match value Y_i, Neter *et al*. (1965) characterise this error by defining a new random variable

$$Y_i^* = \begin{cases} Y_i & \text{with probability } \lambda \\ Y_k & \text{with probability } \gamma\,(k \neq i) \end{cases}$$

where $\lambda + (N-1)\gamma = 1$. This has been dubbed the *exchangeable linkage errors* (ELE) model. Analyses using the linkage error affected variable Y_i^* are unbiased for means, but variances are inflated (increasing type II errors), correlations between Y_i^* and other variables are attenuated, and estimates of regression parameters based on Y_i^* are biased.

A simple extension (Chambers 2009, Kim & Chambers 2012a, b) embeds this model within post-strata assuming no between-stratum matches and 1-1 linkage together with ignorable linkage errors within strata. Given a matrix $\mathbf{X}_q$ of regression covariates and a vector $\mathbf{y}_q^*$ of linked values in stratum *q*, an unbiased estimator of the linear regression of $\mathbf{y}$ on $\mathbf{X}$ under ELE is then

$$\hat{\beta} = \left[\sum_q \mathbf{X}_q^T \mathbf{E}_q \mathbf{X}_q\right]^{-1} \left[\sum_q \mathbf{X}_q^T \mathbf{y}_q^*\right]$$

where $\mathbf{E}_q = (\lambda_q - \gamma_q)\mathbf{I}_q + \gamma_q \mathbf{1}_q \mathbf{1}_q^T$ (with obvious notation). More complex extensions are possible, for example allowing linkage errors only with ‘closer’ units giving a banded diagonal $\mathbf{E}$, and also where λ varies from unit to unit. A maximum likelihood estimator (different to $\hat{\beta}$ above) is also available under additional assumptions on the variances (Chambers 2009, section 2.3).

Other principled ways to analyse linked data have also been suggested, for example Goldstein *et al*. (2012) propose Bayesian methods and multiple imputation.

Analysis of linked data should account for linkage errors, minimally to assess the sensitivity of results, although unbiased analysis quickly becomes challenging even in simple situations. More development and some case studies implementing these methods would be very valuable. In this respect we think that Professor Hand’s challenges 12 and 13 are not ambitious enough, and should be extended to include principled model-based analysis of linked datasets. This would go some way towards statisticians being more specific about the impact of data quality on analytical outputs from administrative data.

References

Chambers, R. (2009). Regression analysis of probability-linked data. *Official Statistics Research Series,* **4**. http://www.statisphere.govt.nz/further-resources-and-info/official-statistics-research/series/volume-4-2009.aspx.

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. doi: 10.1002/sim.5508.

Kim, G. & Chambers, R. (2012a) Regression analysis under probabilistic multi-linkage. *Statistica Neerlandica* **66** 64–79. doi: 10.1111/j.1467-9574.2011.00509.x.

Kim, G. & Chambers, R. (2012b) Regression analysis under incomplete linkage. *Computational Statistics and Data Analysis* **56** 2756–2770. doi: 10.1016/j.csda.2012.02.026.

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.