Statistical methods accounting for human coding habits in databases

Robert L Grant, BayesCamp Ltd, 2018 ©

This is some work I did in 2013 that never led to funding to take it further. I did a lit search around the subject and wrote up ideas, which I share here.

Bottom line: there are none (at least in 2013). I then looked at codes used by GPs (family doctors) in the UK for dementia and incontinence, which I had analysed with colleagues. I found some variation by GP but again there wasn't enough appetite to take it further. I don't have those data or the output of that analysis any more. That's the price of confidentiality.

I also organised a session at RSS conference in Sheffield on "Checking and cleaning in big data". It was not very well attended as everyone had probably gone to hear about something trendier. But the people who were there appreciated the problem, and wanted to learn from experts, and that was pleasing. My invited speakers were Ton de Waal from Statistics Netherlands and Liz Ford from Brighton & Sussex Medical School. You should look them up if you are into this kind of thing. I had someone else from Reuters lined up to talk about automated processing of text streams in real-time but they moved jobs and were contractually gagged, alas.

Anyway, here's the write-up on the preliminary review. You might find it stimulating. I think it's an interesting and under-valued avenue for research. I would have liked to have developed some Bayesian model that incorporated the hierarchical structure of the data by the professional doing the coding, and then included latent variables for coding habits. These could have been developed from a preliminary study to hand-classify coding habits and maybe dimension-reduce them into a manageable number of factors.

Over to you now.

Data linkage

The goal of data linkage is to combine information from different databases into one. When there is not a unique identifying variable for each subject, special techniques have to be employed to find a likely match and unbiased results from any analysis that follows. Established data linkage methods, whether probabilistic or not, typically lead to the creation of a single linked dataset which is then analysed as if it were perfectly matched. This effectively ignores any uncertainty arising from the matching, and can introduce bias if the incorrect matches are different to the correct ones in terms of some of the variables used in the analysis. However, Bayesian approaches by McGlincy ("A Bayesian Record Linkage Methodology for Multiple Imputation of Missing Links", 2004) and Goldstein, Harron and

Wade ("The analysis of record-linked data using multiple imputation with data value priors", 2012) have capitalised on the ease with which computational methods such as MCMC can perform analysis and editing / imputation in a single step. Both approaches allow data to be imputed from conditional distributions if no match is sufficiently probable. Goldstein, Harron and Wade used a multiple imputation approach to create several potential matched datasets in order to capture the uncertainty that arises from the matching process. In none of these papers is there any mention of the possibility of human coding necessitating a multilevel structure to the linkage probabilities and weights.

Automated edit/impute procedures

Large surveys require a computerised approach to checking data for errors and correcting them where possible. A statistical approach can be traced back to Fellegi and Holt's seminal paper ("A Systematic Approach to Automatic Edit and Imputation", 1976). Census agencies, particularly in the USA and the Netherlands, have led the way in developing methods and software, but adoption among a broader statistics community has been rare. De Waal, Pannekoek and Scholtus ("Handbook of statistical data editing", 2011) provide a comprehensive review of edit/impute methods. A number of common forms of human error are detailed but none of the methods incorporate the identity of the individual recording the data, perhaps because national surveys typically do not have more than one record per individual. There is however a passing reference (p. 28) to a certain type of error being made consistently throughout different variables.

The Fellegi-Holt paradigm aims to produce "internally consistent records, not the construction of a data set that possesses certain distributional properties" (de Waal, Pannekoek, Scholtus, p. 63).

de Waal, Pannekoek and Scholtus note that influential and unusual observations are still generally identified by computer and considered by experts, possibly by contacting the source.

Coding bias

Because of the prominence of coding systems in medical data (for example, ICD or Read codes), a search of the Medline database was conducted for the terms "coding bias" (13 retrieved, none relevant), "interviewer bias" (40 retrieved, likewise). These searches were augmented by searches for the same terms on Google Scholar and Google Web Search, and consideration of references in any partly relevant documents.

Jameson & Reed ("Payment by results and coding practice in the National Health Service", 2007) and Joy, Velagala & Akhtar ("Coding: An audit of its accuracy and implications", 2008) suggest that coding can lead to a considerable change in a healthcare provider's income within the British NHS's Payment By Results scheme. This has been emphasised as a system-wide problem by the Audit Commission and NHS Connecting For Health.

Systematic investigation of the bias arising from coding is much rarer. Lindenauer and colleagues ("Association of diagnostic coding with trends in hospitalizations and mortality of patients with pneumonia, 2003-2009.", 2012) conducted a thorough analysis of coding trends over time for hospital patients with pneumonia and/or sepsis, and found that the use of pneumonia codes had declined between 2003 and 2009, while codes for sepsis secondary to pneumonia, and respiratory failure with pneumonia, had increased. While mortality rates (adjusted for age, sex and co-morbidities) in each category had dropped significantly over the same time period, taken together as a single category, the mortality rate had not significantly changed. The authors suggest that patients that would have been at high risk of dying with a pneumonia code in 2003 were increasingly given sepsis or respiratory failure codes (thus artificially improving mortality rates in the pneumonia group), where they became comparatively low-risk patients. Meanwhile, advances in treatment for sepsis had improved mortality in the other two groups' higher-risk patients. Commenting on the medical website Medscape (http://www.medscape.com/viewarticle/765523), Shorr described the coding bias exposed by this study as "not comparing apples with apples and oranges with oranges [but]… mixing things up and making fruit salad".