Disclosure-Protected Regression Coefficients with Linked Micro-Data

Industry Representative:

Daniel Elazar, Australian Bureau of Statistics (ABS)

Problem Moderators:

Tony Pettitt, Queensland University of Technology, Brisbane, Australia
Tapan Rai, University of Technology Sydney, Sydney, Australia

Student Moderator:

Elizabeth Ryan, Queensland University of Technology

Group Participants:

Scott Alexander, University of Technology Sydney
Jacobien Carstens, Royal Melbourne Institute of Technology
Anagi Gamachchi, Royal Melbourne Institute of Technology
Joanne Hall, Queensland University of Technology
Douglan Stebila, Queensland University of Technology

Introduction

The ABS mission is to encourage informed decision making, research and discussion within governments and the community. The ABS releases micro-data to approved users but it has a legal obligation to ensure that “disclosure” is unlikely. “Disclosure” occurs if data collected by the ABS is attributed to the person or business from which it was collected. Confidentialisation procedures are used to ensure that the disclosure risk is acceptably low.

There is growing user demand for analyses involving, for example, demographic details, geographic details and measurements details (e.g., income in dollars). In developing a system to handle these issues, the ABS uses such terminology as the Analyst, the Integrating Authority (IA), and the Data Custodian (DC). The Analyst can request calculations such as means, quantiles and regression estimates. The Remote Server (RS) manages the release of statistical, usually aggregate, output such as a regression analysis. The system deals with data linkage and disclosure risk which is described below.

An IA is trusted to integrate, or link, micro-data, e.g., data about members of a household, collected by two or more government agencies, DCs. An IA's role is to maximize the inherent value of government data for the benefit of society whilst protecting the legislative requirements of all DCs. Balancing utility and disclosure risk is an issue for an IA. Potential analysts of linked micro-data include the DCs, non-custodians (e.g., academics) and the IA.

An example of this system is as follows. DC 1 is the ABS with Census micro-data (e.g., income range, employment status, for household members). DC 2 is the Department of Immigration and Citizenship (DIAC) with micro-data obtained from the Settlements Database (e.g., visa class). The IA is the ABS and links micro-data using names and addresses. The Analyst is the DIAC and is interested in employment outcomes by visa class.

Such an analysis could be carried out using a variable from one DC, e.g., visa class, and a variable, household income, from another DC.

The ABS **problem** involves designing the analysis carried out by the RS such that statistical analyses requested by an Analyst are carried out with minimal risk of disclosure but with as little perturbation from the true results. The risk of disclosure is controlled by perturbing analyses.

MISG 2013 Contribution

The group's first activity was to familiarise itself with the ABS system. The problem was restricted to regression analysis and the group's need to fully understand the nature of this problem.

Although several different types of regression are possible, it was decided to concentrate on the Linear Model for which disclosure might be affected more readily than non-linear models. Additionally, a binary response was considered as it offers simplification of analysis for which disclosure might be affected more readily than counts, continuous data or coarsened continuous data.

The group considered a simplified problem that may be sufficiently rich to indicate effective confidentialisation strategies. The simplified problem is as follows.

- DC 1 has a y value, the response or dependent variable, e.g., employment, for each case and was matched with DC 2 data.
- DC 2 has a set of K, x-variables or covariates, e.g., gender, tertiary qualification ($25 \leq$ age < 34), for a large number of cases or individuals, e.g., $n = 20,000$.
- The RS combines microdata and carries out regressions requested by DC 2 and provides DC 2 with regression estimates and standard errors for requested models and additional summaries.
- The designer of the RS has to reduce the risk of disclosure to DC 2 of DC 1's data.
- The RS must have some automatic perturbations of analyses to reduce the risk of disclosure when a succession of regressions of y on a set of xs is requested by DC 2.
- To simplify further, the RS provides estimates of regression parameters by using ordinary least squares. There is an explicit solution and well known theory based on matrix algebra.
- Minimization of the risk of disclose involves data perturbations or deletions that must not change regression estimates nor standard errors by a substantial amount from values based on the original data set.

We considered *Entropy based decisions about disclosure.* Entropy is a measure of uncertainty. For binary Y the maximum entropy is log(2) or 1 bit. A procedure for disclosure risk minimization was discussed.

We discussed ϵ Differential Privacy (see Dwork (2006), Zhang et al (2012)). It was found that the theory in Zhang et al (2012) was in error so we proposed to develop a new definition of ϵ Differential Privacy that involved absolute differences not multiplicative and corrects the error of Zhang et al (2012).

We discussed the notion of "Split Leverage" as a risk of disclosure. If a data set has a leverage point, then it is possible to carry out an analysis to recover the data value for this point. We developed a method for obtaining an approximation to the regression line of the whole set, by splitting the set into subsets, none of which contains a leverage point.

References

Dwork, C. (2006). Differential Privacy, in The 33rd International Colloquium on Automata, Languages and Programming, New York: Springer. pp 1- 12.

Zhang, J., Zhang, Z, Xioa, X., Winslett, M (2012) Functional Mechanism: Regression Analysis under Differential Privacy. Proceedings of the VLDB Endowment (PVLDB), Vol. 5, No. 11, pp. 1364-1375