IAB-DiskAB

Remote data access and the risk of disclosure from linear regression: An empirical study

Philipp Bleninger (KEM)

In the endeavor of finding ways for easy data access for researchers not employed at a statistical agency remote data access seems to be an attractive alternative to the current standard of either altering the data substantially before release or allowing access only at designated data archives or research data centers. Data perturbation is often not accepted by the researchers since they do not trust the results from the altered data sets. But on-site access puts some heavy burdens on the researcher and the data providing agency both in terms of time and money. Remote data access or remote analysis servers that allow to submit queries without actually seeing the micro data have the potential of overcoming both these disadvantages. However, even if the micro data is not available to the researcher directly, disclosure of sensitive information for individual survey respondents is still possible.

In this paper we illustrate how an intruder could use some commonly available background information to reveal sensitive information using simple linear regression. We demonstrate the real risks from this approach with an empirical evaluation based on the IAB Establishment Panel.

Although this kind of attacks can easily be prevented once the agency is aware of the problem, this small simulation aims to emphasize that there might be many ways to obtain sensitive information using multivariate analysis and not all of them are obvious. Thus, agencies thinking about actually implementing some form of remote data access should consider carefully which queries could be allowed by the system.

Dienstag, 14. September 2010

**14.00 Uhr**

Sitzungssaal **126a**