記述統計量に基づく秘匿すべき回帰モデルの検証

一橋大学経済研究所　　白川清美

オンサイト施設における公的統計調査ミクロデータの利用では、研究成果が公開可能か否かの持ち出し審査がある。この審査は、ESS Net（A Network of Excellence in the European statistical System in the field）のガイドラインに基づいているが、これらの基準の根拠が示されていない。それゆえ、Shirakawa et al [1]は、平均、分散、歪度および尖度などの記述統計量に掛かる検証をした。ただし、回帰モデルまでは検証していない。

そこで、本研究では SDC（Statistical　Disclosure Control）の視点に基づき、回帰モデルの計算過程において算出される平均、分散および共分散などの記述統計量による安全か否かの検証を行う。一般的に、統計解析では、統計モデルの当てはまりの良さを AIC や BIC で評価する。しかしながら、これらの指標は回帰式の説明力を示してはいるものの、個票データの数値を復元することは難しい。したがって、式の当てはまりとは異なる基準で持ち出し審査を考える必要がある。

図１の単回帰モデル $\hat{y} = a + bx$ の最小二乗推定量は以下の式で求めることができる。

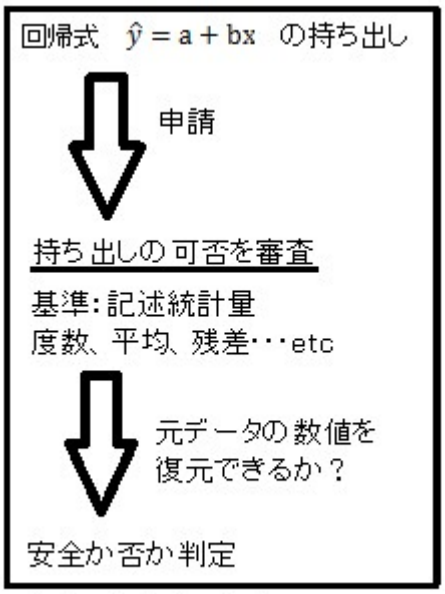

図1　持ち出し審査のイメージ

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad \cdots(1)$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \qquad \cdots(2)$$

上記（1）式より、回帰係数$\hat{b}$は x と y の共分散と分散の比である。分散は x と y の残差平方和と度数から計算できる。残差平方和の計算には残差が必要であり、残差の計算には平均が必要である。このことから、特定の記述統計量から元の数値が復元できる可能性に着目すればよいことが分かる。なお、平均は公表されることが多いので、残差あるいは残差平方が分かれば、元の数値が復元することができる。よって、残差および残差平方は公表できない。

詳細については当日報告する。

参考文献

[1] Empirical Analysis of Sensitivity Rules: Cells with Frequency Exceeding 10 that Should Be Suppressed Based on Descriptive Statistics（共著）Privacy in Statistical Databases, UNESCO Chair in Data Privacy, International Conference, PSD 2016, Dubrovnik, Croatia, September 14-16, 2016, Proceedings 9867 of the series Lecture Notes in Computer Science 巻 28-40 頁 2016 年 学術雑誌 ISBN 978-3-319-45380-4