

決定木と Chow 検定によるリコーディング境界値の分析

独立行政法人統計センター
一橋大学経済研究所

阿部 穂日
白川 清美、千葉 亮太

公的統計調査の個票データから匿名データを作成する際には、開示リスクを軽減するため、様々な統計的開示抑制（SDC: Statistical Disclosure Control）を講じている。その一つとして、トップコーディングとボトムコーディングがある。これは、各変数に対し上限値（または下限値）を設定し、その境界値以上（または以下）のデータを一つの階級としてまとめる手法である。これにより、極端に大きな（または小さな）値のデータを秘匿することができる。

しかしその一方で、上限値（または下限値）による情報損失が問題となる。例えば実際のデータが複数の分布から成る場合、境界値を分布の境界に設定すると、秘匿された分布を公開されたデータから推測することが出来なくなる。

そこで、本研究では、トップコーディングとボトムコーディングの境界値検証のため、決定木と Chow 検定を組み合わせることで、適切な境界値を探索する手法について、シミュレーションデータを用いて考察した。

詳細な結果については、報告当日に発表する。

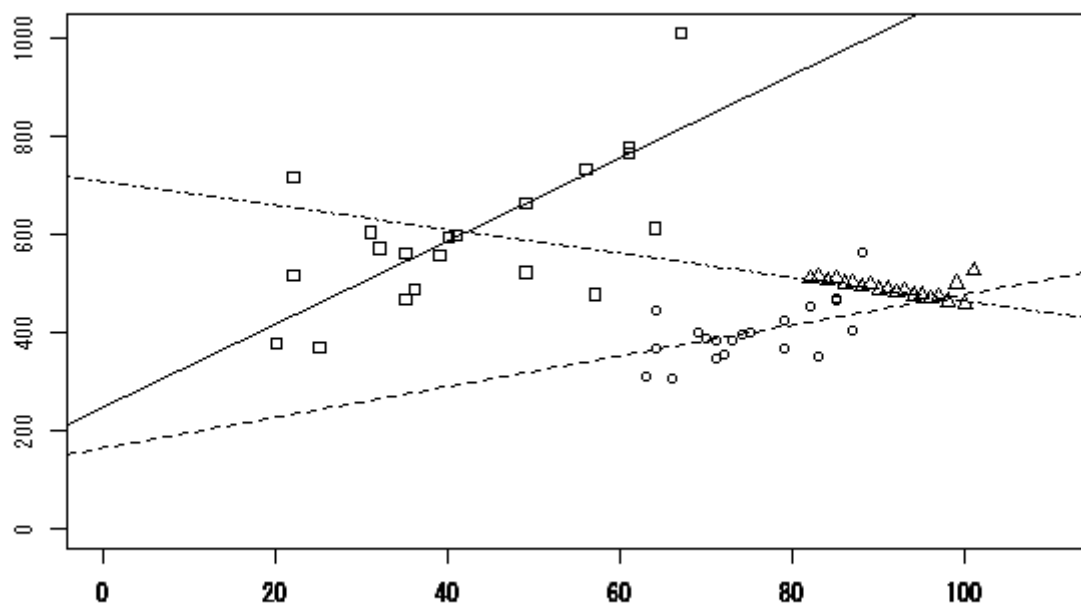


図 1 シミュレーションデータを用いた決定木による層分けの例

【参考文献】

Shirakawa, Kiyomi. "Multi-Stratification for Outlier Detection based on the Graphical Model: Evaluation by Chow Test and AIC." http://www.nstac.go.jp/services/society_paper/26_05_02.pdf. 2017 年 6 月 30 日アクセス。

阿部穂日、白川清美、千葉亮太。「匿名データの作成のためのリコーディング —決定木による最適な境界値の分析—」, 2016 年度統計関連学会連合大会, 2016 年 9 月 6 日。