DSSV 2023 Data Science, Statistics & Visualisation
ECDA 2023 European Conference on Data Analysis

The effect of data aggregation on ordinary least-squares estimation and model selection procedures

P.C. Schoonees[1] and N.J. le Roux[2]

[1]Erasmus University Rotterdam, Econometric Institute, Burgemeester Oudlaan 50, 3062 PA Rotterdam, Netherlands
[2]Stellenbosch University, MuViSU (Centre for Multi-Dimensional Data Visualisation), Department of Statistics and Actuarial Science, Stellenbosch 7600, South Africa

The study of the effect of aggregation on statistical models has a long history in econometrics [see 1, for example]. Aggregating over individuals belonging to different groups before performing linear regression is known to induce a so-called aggregation bias in the ordinary least-squares (OLS) coefficient estimates compared to those obtained without aggregation. There are however situations where this problem is unavoidable, such as when data from different views are merged. For example, neuroscientific studies may combine a sample of individuals who supply EEG data on a set of stimuli (i.e., the groups) with data which reports the aggregated behaviour of a second sample of individuals.

Assume a matrixvariate normal data matrix $\mathbf{V} = \begin{bmatrix}\boldsymbol{Y} & \mathbf{X}\end{bmatrix}$, where $\boldsymbol{Y}$ and $\mathbf{X}$ denote the response vector and matrix of independent variables, respectively [2]. Denote by $\mathbf{G}$ an indicator matrix which assigns each row of $\mathbf{V}$ to one of K groups. The rows of $\mathbf{V}$ are assumed to be independent but not identically distributed since each of the groups have a different mean matrix, whereas a common covariance matrix $\boldsymbol{\Sigma}$ is assumed for the columns. The aggregated data then consists of the data matrix $\mathbf{G}\left(\mathbf{G}^\top\mathbf{G}\right)^{-1}\mathbf{G}^\top\mathbf{V} = \overline{\mathbf{V}}$, which contains the group-wise averages as rows.

In this work, we study the effect such linear aggregations have on model estimates and model selection procedures. Relying on the distributional assumptions introduced above, we report an expression for the bias of the maximum likelihood estimator of the covariance matrix when $\overline{\mathbf{V}}$ is substituted for $\mathbf{V}$. Moreover, it is noted that theoretically the non-intercept OLS estimates derived from $\overline{\mathbf{V}}$ are statistically independent of those derived from $\mathbf{V}$. This is illustrated with a simple simulation experiment.

The adverse effects of aggregation on common model selection procedures are investigated in a second simulation study. Situations are investigated where not all explanatory variables may contain information about the outcome variable. The findings and implications are briefly discussed.

Keywords— aggregation bias, matrixvariate normal, ordinary least-squares, model selection.

[1] H. Theil, *Linear aggregation of economic relations*. North-Holland Publishing Company, Amsterdam, 1954.
[2] A. K. Gupta and D. K. Nagar, *Matrix variate distributions*. Chapman and Hall/CRC, 2000.