MR4718776 62G08 62R07

Dai, Chi-Shian (1-WI-S)**; Shao, Jun** (1-WI-S)

Kernel regression utilizing heterogeneous datasets. (English. English summary)

Stat. Theory Relat. Fields **8** (2024), *no.* 1, 51–68.

This article is concerned with kernel regression estimation when two cross-sectional datasets, namely, the primary (or internal) and secondary (or external) ones, are combined. Throughout, it is assumed that the internal dataset contains a response $Y \in \mathbb{R}$ and a vector of covariates $\mathbf{U} \in \mathbb{R}^p$. The main focus of this study is on how the external dataset can improve efficiency in nonparametric kernel estimation for the conditional expectation of Y given $\mathbf{U}$ under the internal population, $\mu_1(\mathbf{u}) = E(Y|\mathbf{U} = \mathbf{u}, D = 1)$, where $D = 1$ indicates internal population. The local constant or Nadaraya-Watson regression estimator is chosen as the estimation method. A novel feature of this study is dealing with the case in which two datasets are drawn from different populations. Such heterogeneity in populations arises, for instance, when two datasets are collected in different ways and/or different time periods.

The analysis starts from the case in which the external dataset also contains $(Y, \mathbf{U})$. Then, depending on three scenarios, three estimators for $\mu_1(\mathbf{u})$ are proposed. To be more concrete, the estimators $\widehat{\mu}_1^{E_j}(\mathbf{u})$ for $j = 1, 2, 3$ correspond to the cases in which (i) the external dataset is from the same population, (ii) the external dataset is from a different population, and (iii) the external dataset from a different population possesses additional information, respectively.

The analysis is further extended to the case in which the external dataset instead contains $(Y, \mathbf{X})$, where $\mathbf{X} \in \mathbb{R}^q$ is a sub-vector of $\mathbf{U}$ with $q < p$. Then, depending on the above three scenarios, three estimators $\widehat{\mu}_1^{C_j}(\mathbf{u})$ for $j = 1, 2, 3$ are additionally proposed. All these estimators can be obtained through a constrained optimization using the external information as a constraint.

Pointwise asymptotic normality of $\widehat{\mu}_1^{E_j}(\mathbf{u})$ and $\widehat{\mu}_1^{C_j}(\mathbf{u})$ for a given $\mathbf{u}$ is demonstrated under some regularity conditions. The simulation study also confirms superiority of $\widehat{\mu}_1^{E_3}(\mathbf{u})$ (resp., $\widehat{\mu}_1^{C_3}(\mathbf{u})$) over $\widehat{\mu}_1^{E_1}(\mathbf{u})$ and $\widehat{\mu}_1^{E_2}(\mathbf{u})$ (resp., $\widehat{\mu}_1^{C_1}(\mathbf{u})$ and $\widehat{\mu}_1^{C_2}(\mathbf{u})$) when the populations of two datasets differ and additional information from the external population is available. *Masayuki Hirukawa*