K-Nearest Neighbor (kNN) Imputation

Presented by Geoffrey Hubona

K-Nearest Neighbor (kNN) imputation

- ***Imputation*** is a procedure used to fill in missing values by using substitutes
- **Nearest Neighbor (NN)** imputation is motivated by the idea that records, or observations, characterized by similar **X-variable** values would be characterized by similar **Y-variable** values
- ***k Nearest Neighbor*** **(kNN)** imputation uses values observed for ‘k’ reference records (1 or more) that have characteristics similar to the missing target records.

K-Nearest Neighbor (kNN) imputation

- **kNN** imputation can use either one single neighbor (when **k = 1**) as the donor for the missing **Y-variables** of the target records or a simple or weighted average of **k > 1** near neighbors to fill in the **Y-variables.**
    - The weights may be chosen to reflect the degree of similarity in the ***X-variables***.
- **kNN** imputation methods are **non-parametric** or **distribution-free** in that they do not rely on any underlying distribution for estimation.

Distance Metrics

- **NN *imputation methods*** use different distance metrics to determine the similarity between **target** (missing Y-variable values) and **reference** records (complete x- and Y-variable values).
- Absolute distances are often used, for example **Euclidean** or **Mahalanobis distance functions**:

$$d_{ij} = \sum_{l=1}^{p} c_l |x_{il} - x_{jl}| \qquad (1)$$

where x_{il} is the value of the X-variable l for target record i, x_{jl} is the value of the X-variable l for reference record j, p is the number of X-variables, and c_l is the coefficient for variable x_l.

Quadratic Distance Metrics

- However, with **kNN *imputation methods*** quadratic distances are also often used:

$$d_{ij}^2 = (x_i - x_j)W(x_i - x_j)' \qquad (2)$$

where x_i is the $(1 \times p)$ vector of x-variables for the ith target record, x_j is the $(1 \times p)$ vector of x-variables for the jth reference record, and W is a $(p \times p)$ symmetric matrix of weights.

kNN Distance Metrics

- For the ***squared Euclidean distance,*** the weight matrix, **W**, is the diagonal identity matrix, giving equal weight to each **X-variable**.
- The ***squared Euclidean distance*** gives more emphasis to larger differences than the absolute difference distance (eq. 1) because the differences are squared.
- The ***Mahalanobis distance*** is produced by using the inverse covariance matrix of the **X-variables** for **W**.

Number of Neighbors (k)

- LeMay and Temesgen (2005) compared the use of the nearest neighbor, the average of three near neighbors and the distance-weighted average of three near neighbors.
  - The reported that the estimates may not be within the bounds of reality if more than one neighbor is used.
  - The **complex variance-covariance structure** and the natural possibilities of the Y-variable are retained only when **k = 1**.
- When **k = 1** all variability in the observations is preserved; when **k > 1** smoothing occurs as estimates are based on averages of multiple observations.

Number of Neighbors (k)

- With small **k** values, **NN** methods may produce less accurate results than using the mean over all observations for every prediction.
    - Accuracy of estimates improves with increasing **k** to an optimal choice of **k**.
    - With a larger number of reference records, larger values of **k** can be applied.
- The optimal choice of **k**, and the distance metric including weights, and **X-variables** is difficult to determine.

Best Methods

- The choice of **X-** and **Y-variables**, the ***distance metric*** and **k** all contribute to the imputation error.
- The optimal choice of **k**, and the distance metric including weights, and **X-variables** is difficult to determine.
- Differences in ***data structure***, ***selection of Y-variables*** and ***availability of X-variables*** suggest that no single choice of distance metric, **X-** and **Y-variables** and **k** gives the best results for all applications.
    - These choices are best decided case-by-case.