3. Data Pre-processing

Statistical Modelling & Machine Learning

Jaejik Kim

Department of Statistics
Sungkyunkwan University

STA3036

Data preparation

- Data preparation can make or break a model's predictive ability.

(1) Raw data: Records in a database.

User	Whom	Dialed/Received	Call start time	Call end time	⋯
User1	User37	Dialed	09:42:13	09:51:10	⋯
User2	User25	Received	10:11:10	10:11:42	⋯
User1	User15	Dialed	10:13:25	10:37:35	⋯
User3	User86	Dialed	10:17:02	10:18:26	⋯
⋮	⋮	⋮	⋮	⋮	⋮

(2) Define Objects for the analysis.

(3) Create important predictors from raw data (Encoding of predictors).

(4) Missing values & outliers.

(5) Data transformation.

Feature Engineering

- Feature engineering: How the input variables are encoded.
  - Domain knowledge is required.
  - Combinations of predictors can sometimes be more effective than individual predictors (e.g., ratio of two inputs).
  - Type of predictor: Different models have different sensitivities to the type of input variables in the model.
  - Consider your model and true relationships with the output.

E.g., Ways to create a predictor from time of customer's credit card transaction.

- Database record: 18:42:32, 10/09/2020 Friday.
- Isolating the month, year, and day of the week as separate predictor.
- The number of days since the previous transaction.
- Day or night time.
- Spring / summer / fall / winter
- Weekday / Weekend / Holiday.

Transformation of Inputs

- Why transformation?
  - Some model may have strict assumptions or requirements (e.g., LDA require normally distributed predictors).
  - Specific characteristics of data (e.g., outliers).
  - For computational efficiency of model estimation (e.g., neural networks require normalized inputs for the convergence of the algorithm).

- Centering & Scaling:
  - Centering: Centered input variable with mean zero.
  - Scaling: Remove sale effect.
  - Standardization: Centering + Scaling (e.g., regularization methods).
  - $[0, 1]$ scaling: All values between 0 and 1 (transformation using min. & max. values).

  $\Rightarrow$ Drawback: Loss of interpretability of the individual values.

Transformation for Skewness

- Input variables with highly skewed distributions might lead to poor prediction.
  - In fact, removing the skewness does not guarantee the improvement of models.
  - Models with normality assumption (e.g., LDA, QDA) or polynomial calculation (e.g., polynomial regression, support vector machine, etc.) for inputs could have harmful effects due to inputs with skewed distribution.
  - Polynomial calculation can be dominated by the tails of skewed distribution.
- Sample skewness:

$$skewness = \frac{\sum(x_i - \bar{x})^3}{(n-1)s^{3/2}},$$

where $\bar{x}$ and s is the sample mean and standard deviation, respectively.

  - $skewness \approx 0 \Rightarrow$ Roughly symmetric.
  - $skewness > (<)\ 0 \Rightarrow$ Right (Left) skewed

Transformation for Skewness

- Box-Cox transformation for $x > 0$:

$$x^{(\lambda)} = \begin{cases} \frac{x^\lambda - 1}{\lambda}, & if\ \lambda \neq 0 \\ \log(x) & if\ \lambda = 0 \end{cases}$$

  - $\lambda = 2$ (square transformation), $\lambda = 0.5$ (square root), $\lambda = -1$ (Inverse).

- Box-Cox transformation for $x > -\delta$:

$$x^{(\lambda)} = \begin{cases} \frac{(x+\delta)^\lambda - 1}{\lambda}, & if\ \lambda \neq 0 \\ \log(x+\delta) & if\ \lambda = 0 \end{cases}$$

Transformation for Skewness

- Yeo-Johnson transformation for $-\infty < x < \infty$.

$$x^{(\lambda)} = \begin{cases} \frac{(x+1)^{\lambda}-1}{\lambda}, & \textit{if } \lambda \neq 0,\ x \geq 0 \\ \log(x+1) & \textit{if } \lambda = 0,\ x \geq 0 \\ -\frac{(-x+1)^{(2-\lambda)}-1}{2-\lambda} & \textit{if } \lambda \neq 2,\ x < 0 \\ -\log(-x+1) & \textit{if } \lambda = 2,\ x < 0 \end{cases}$$

- In the Box-Cox and Yeo-Johnson transformations, λ is estimated by maximizing the profile normal likelihood function.

Transformation for Data Bounded in $[0, 1]$

- Problems of output & Inputs bounded in $[0, 1]$:
  - Proportion data.
  - Output: The range of predicted values can be out of the original range.
  - Inputs: Normality assumption.

- Logistic transformation: $[0, 1] \rightarrow (-\infty, \infty)$.

$$x^* = \log\left(\frac{x}{1-x}\right), \ 0 < x < 1.$$

- When proportion data have many zero values, the logistic transformation returns many $-\infty$ values. In that case, arcsine transformation can be considered.

- Arcsine transformation: $[0, 1] \rightarrow (0, \pi/2)$.

$$x^* = \arcsin(\sqrt{x}), \ 0 < x < 1.$$

Transformation for Time or Sequential Data

- Time output or inputs: Data measured at discrete time points.
- Sequential output or inputs: Data measured with sequence.
- Noise or outliers could blur patterns.
- Smoothing techniques for reducing noise or outlier effects:
    - Moving average: The window size is important. Too large window eliminates patterns and too small window cannot reduce noise.
    - Running median: It can reduce significant outlier effects.
    - Smoothing splines.

Transformation for Time or Sequential Data

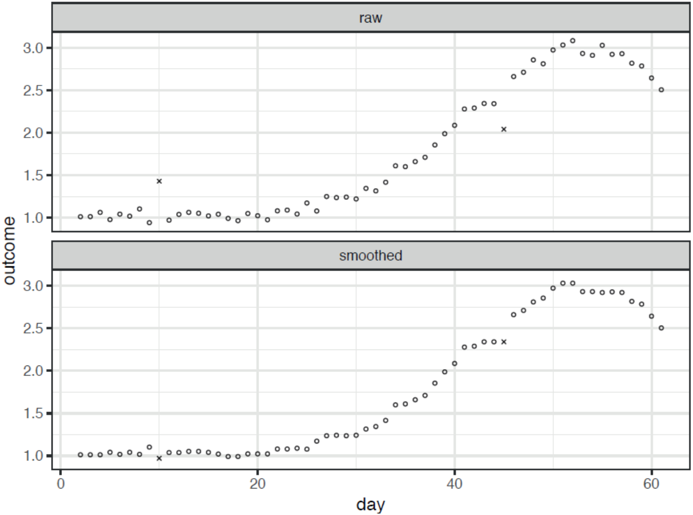


Figure: Results of a 3-point running median.

Discretization of Continuous Output or Inputs

- Discretization (or Binning): Transformation of a continuous variable into a categorical variable.

- Reasons for discretization:
  - Better interpretation (e.g., age of salary men → 40 years old or not).
  - To avoid the problem specifying functions of Y and X.
  - Reduction of the data variation.

- Cut-points for discretization:
  - User-driven cut-points.
  - Percentile (e.g., quartile values, median, etc.)
  - Output: Cut-points that gives the best performance.
  - To avoid overfitting, determination of cut-points must be performed within the resampling process.

- Caution: Discetization of continuous inputs should be the last method.

Outliers

- It is hard to define outliers for given dataset.
- Check points for outliers:
  - Data values should be scientifically valid (e.g., positive salary).
  - Data recording error.
  - Very large or small values due to skewed distributions.
  - Data collected from a different population due to the data collection procedure.
- Some models robust to outliers:
  - Tree-based classification model.
  - Support vector machine.
  - Robust estimation methods (e.g., Huber estimation, M-estimation, etc.)

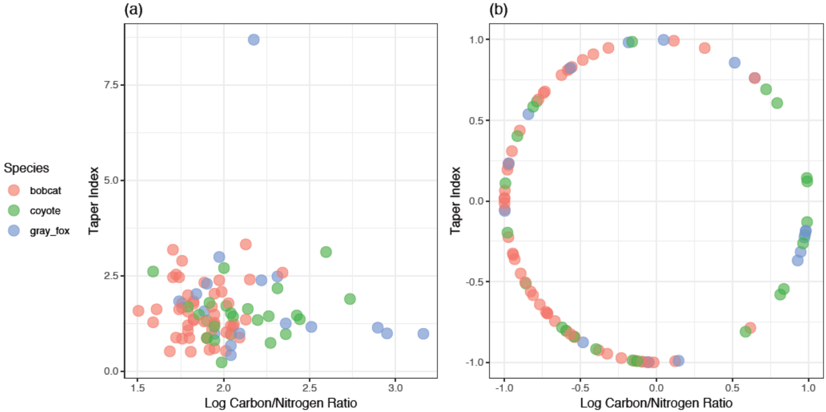
Spatial Sign Transformation

- Spatial sign: All data points are mapped on the multidimensional sphere. Transformation for a set of variables.

$$x_{ij}^* = \frac{x_{ij}}{\sqrt{\sum_{j=1}^{p} x_{ij}^2}}.$$

  - Standardization is required prior to the spatial sign.
  - Highly skewed variables should be also transformed into symmetric data.
  - it is also referred to as global contrast normalization and often used for image data.
  - It can be considered to avoid the damage of extreme outliers.
  - After the spatial sign transformation, variable selection is not possible.

Spatial Sign Transformation



Missing Data

- The small number of missing values relatively to the sample size ⇒ Deletion of all obs. with missing values.
- Not ignorable size of missing values ⇒ Poor prediction due to the small size of training data.
- Efforts for imputing missing values are required.
- To impute missing values, we have to understand the mechanism that missing values occur in the data.
- Imputation method by the mechanism that missing values occur.

Types of Missing Data

Types of missing data:

1. Missing completely at random (MCAR):
   - The likelihood of a missing result is the same for all data points (observed or unobserved).
   - This means that whether something is missing is not related to the subject of the missing data.
   - E.g., a questionnaire might be lost in the post, or a blood sample might be accidentally damaged in the lab.

2. Missing at random (MAR):
   - The likelihood of a missing result is not equal for all data points (observed or unobserved).
   - The prob. of a missing result depends on the observed data, but not on the unobserved data.
   - E.g., In depression survey, males are less likely to provide their response. The likelihood of responding is indep. of their actual depression level. After accounting for gender, these data can still induce parameter bias.

Types of Missing Data

3. Missing not at random (MNAR):
    - The tendency for a variable to be missing is a function of data that are not available.
    - E.g., Persons with very low or very high income levels are less likely to provide their personal income. There are no information related to their income in the dataset.

- Properties of each type of missing data:
    - MCAR & MAR: Ignorable non-response.
    - MNAR: Nonignorable non-response.
    - In practice, it is hard to distinguish MCAR and MAR.
    - It is not easy to impute missing values for MNAR. Most imputation methods assume at least MAR.
    - Even though MAR assumption is not testable, it may hold approximately if enough variables are included in the imputation methods.

Methods to Handle Missing Data

1. Deletion methods: MCAR is assumed.
   - Listwise deletion (complete case analysis): Delete all incomplete cases (obs. with missing values) ⇒ Loss of prediction power and bias (not MCAR cases).
   - Pairwise deletion (available case analysis): Incomplete cases are deleted on an analysis-by-analysis basis. ⇒ Hard to compute standard error due to the change of sample size.

2. Imputation methods:
   - Single imputation: A missing value is replaced with a value from a imputation method ⇒ Imputation uncertainty exists.
   - Multiple imputation: The imputation process is repeated multiple times resulting in multiple imputed datasets ⇒ Imputation uncertainty is accounted for.

3. Likelihood-based methods: No imputation. The missing data is handled within the analysis model.
   - EM algorithm.
   - Full information maximum likelihood estimation.

Before Handling Missing Data

Characterization for patterns of missingness using exploratory data analysis.

- Models such as logistic regression or recursive partitioning can be used to predict the prob. of missingness (i.e., observed and unobserved groups for a given variable).
- Before imputation, it is important to describe how variables are simultaneously missing.
  - Cluster analysis to identify groups of variables that tends to be missing on the same obs.
  - Variables in the same cluster cannot be used for imputation.
- Exploration of the distribution of non-missing Y by the number of missing variables in X.

Missing Values for Y

- It is common to discard obs. having missing Y.
- Before discarding such obs., it is necessary to investigate missing pattern of Y.
- To check whether missing of Y has MCAR, logistic regression or recursive partitioning can be performed for missing and non-missing Y groups.
  - Significant inputs or good fitting $\Rightarrow$ Not MCAR $\Rightarrow$ Use Y for imputation of missing values of X's $\Rightarrow$ After the imputation of X's, discard obs. having missing Y.
- Not MCAR for categorical Y $\Rightarrow$ sometimes missing Y might be considered as its own category. (NOTE: This approach is not valid for X's even when MCAR holds).

Imputation Methods

- Mean or Median imputation: When the input variable is not related to all other inputs, it may be substituted for missing values without much loss of efficiency.

- Conditional mean: When the input with missing values are related to the other inputs ⇒ Predictive models for X with missing values.

- K-Nearest Neighbors:
  - Find k closest obs. to the obs. with a missing value from complete cases ⇒ Mean, median, or mode of the k closest obs.
  - $k = 5 \sim 10$ is commonly used.
  - The Gower similarity can be used for both continuous and categorical inputs.
  - Gower similarity: For categorical X_j, if both obs. have the same category value, it has 1. Otherwise 0. For continuous X_j, compute $1 - \frac{|x_{ij} - x_{i'j}|}{R_{X_j}}$, where R_{X_j} is the range of X_j.

Imputation Methods

Predictive mean matching (PMM):

- Replace a missing value for X_j being imputed with the actual value from the donor obs. (usually complete cases).
    - It builds a model for X_j using non-missing obs. (nonlinear models are possible).
    - Then, it finds a complete obs. (donor) whose predicted value are closest to the predicted value of the missing value.
- It needs to avoid overuse of good donors to disallow excessive ties in imputed data.

Imputation Methods

Predictive mean matching (PMM):

- To avoid this, it can consider to sample a X_j value from complete obs. by kernel weighted probabilities of the predicted values and the missing target.
- As the distance between the predicted value of complete obs. and missing target increases, the kernel weighted probability decreases.
- Multiple imputation is possible through this sampling procedure.
- No distributional assumptions.

Imputation Methods

- Tree-based models:
  - Tree can be constructed in the presence of other missing data.
  - For each X_j with missing values, tree model should be constructed.
  - Single tree model: Low bias and high variance, but fast.
  - Random forests: Reduction of variance, but slow. If a lot of X's have missing values, it requires to construct a lot of random forest models.
- Regression models:
  - For complete cases, construct appropriate regression models for X with missing values and the other variables.
  - For categorical inputs with missing values, logistic or multinomial logistic regression models can be used.
  - Replace missing values with predicted value from the model.

Multiple Imputation

- Unless uncertainty of imputation is taken into account, usual standard error and other statistics are invalid.
- This uncertainty would be very complex.
- An easy way to account this uncertainty is parametric or nonparametric bootstrap.
- Multiple Imputation: Random sample from the conditional distribution of the variable with missing values.

Multiple Imputation

- Multiple imputation procedure:
  - E.g., When regression model is used for imputation, imputed values x^* can be generated by $x^* = \hat{x} + r$, where $\hat{x}$ is the predicted value from the regression model and r is a random number from the estimated error distribution of the model.
  - If the regression model does not assume distributions, r is bootstrpped from the observed residuals.
  - Once a complete set is created by an imputation model, the predictive model for Y is applied to the complete set.
  - This procedure is iterated many times.
  - The final prediction model can be constructed by model averaging.

Multiple Imputation

- Estimated variance of an estimator $\hat{\theta}$ from the multiple imputation with M iterations:

$$\hat{Var}(\hat{\theta}) = Var_{within} + \frac{M+1}{M} Var_{between},$$

  - $Var_{within} = \left\{ \sum_{m=1}^{M} \hat{Var}(\hat{\theta}_m) \right\} / M$, where $\hat{Var}(\hat{\theta}_m)$ is the estimated variance of $\hat{\theta}$ from the mth generated complete set.
  - $Var_{between} = \left\{ \sum_{m=1}^{M} (\hat{\theta}_m - \bar{\hat{\theta}})^2 \right\} / (M-1)$.

MICE: Multivariate Imputation by Chained Equations

- MICE employs the idea of the Gibbs sampler.

- Gibbs sampler: Construction of the joint distribution by sampling from conditional distributions iteratively.

- Let $\boldsymbol{X}^{(t)} = (X_1^{(t)}, \ldots, X_p^{(t)})^\top$ be $\boldsymbol{X}$ at iteration t. The Gibbs sampler is as follows:

$$
\begin{array}{rcl}
X_1^{(t)} & \sim & p(X_1 | X_2^{(t-1)}, \ldots, X_p^{(t-1)}) \\
X_2^{(t)} & \sim & p(X_2 | X_1^{(t)}, X_3^{(t-1)}, \ldots, X_p^{(t-1)}) \\
 & \vdots & \\
X_p^{(t)} & \sim & p(X_p | X_1^{(t)}, X_2^{(t)}, \ldots, X_{p-1}^{(t)}) \\
\Rightarrow \boldsymbol{X}^{(t)} & \sim & p(X_1, \ldots, X_p), \ t = 1, \ldots, M.
\end{array}
$$

MICE: Multivariate Imputation by Chained Equations

Multivariate Imputation by Chained Equations (MICE):

- It starts with initial imputations by sampling simply from observed marginal distribution of X_j.
- It constructs the conditional distribution from a predictive model for X_j with missing values using other X variables, and then it samples imputed values from the conditional distribution.
- This procedure is iterated until the distributions of imputed values are converged (usually, 10-20 iterations).
- Multiple chains of imputed datasets $\Rightarrow$ Multiple imputation.

MICE: Multivariate Imputation by Chained Equations

- Let $X_j^{(t)} = (X_j^{obs}, X_j^{*(t)})$ be the jth imputed variable at iteration t, where X_j^{obs} is the observed variable and $X_j^{*(t)}$ is the imputed variable.
- $\theta_j^{*(t)}$: Parameters of the jth conditional model.
- Initial imputation: $X_j^{*(0)} \sim p(X_j),\ j = 1, \ldots, p$.
- Algorithm: This below is iterated until the chain is converged.

$$
\begin{aligned}
\theta_1^{*(t)} &\sim p(\theta_1 | X_1^{obs}, X_2^{(t-1)}, \ldots, X_p^{(t-1)}) \\
\boldsymbol{X}_1^{*(t)} &\sim p(Y_1 | X_1^{obs}, X_2^{(t-1)}, \ldots, X_p^{(t-1)}, \theta_1^{*(t)}) \\
&\vdots \\
\theta_p^{*(t)} &\sim p(\theta_p | X_p^{obs}, X_1^{(t)}, \ldots, X_{p-1}^{(t)}) \\
\boldsymbol{X}_p^{*(t)} &\sim p(Y_1 | X_p^{obs}, X_1^{(t)}, \ldots, X_{p-1}^{(t)}, \theta_p^{*(t)}).
\end{aligned}
$$

MICE: Multivariate Imputation by Chained Equations

Diagnostics for MICE algorithm:

- Convergence:
  - Sequences (streams) of imputed values have no trend and they are intermingled.
  - The variance between sequences is not larger than the variance of individual sequences.
- Checking the imputations:
  - For each imputed variable, compare density estimates for observed data and imputed data.
  - These density estimates should be similar.

MICE: Multivariate Imputation by Chained Equations

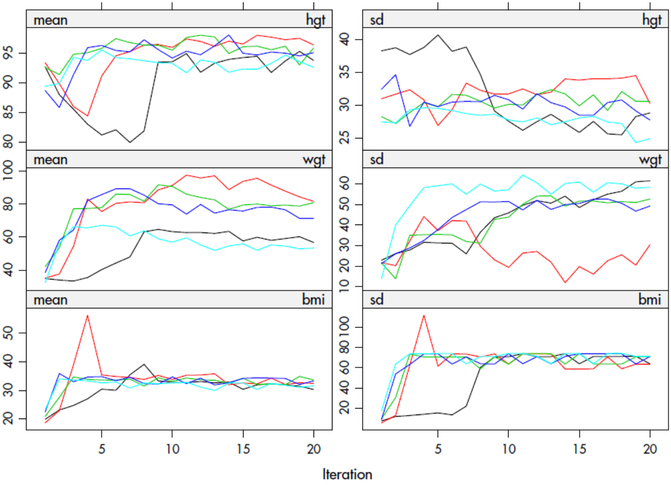


Figure: Non-convergence case.

MICE: Multivariate Imputation by Chained Equations

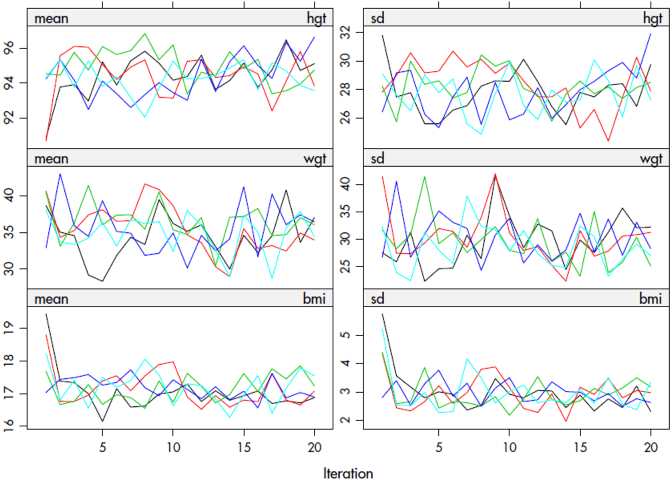


Figure: Convergence case.

MICE: Multivariate Imputation by Chained Equations

Considerations for MICE:

- Choice of imputation model: Predictive imputation models for continuous or categorical variable.
- Selection of predictors in the predictive imputation model and transformation of predictors.
- Order in which variables is imputed: It does not matter if the number of iterations is enough. However, convergence speed may depends on the order.
- The number of imputed datasets: It should be enough to access uncertainties of imputation.

MICE: Multivariate Imputation by Chained Equations

Problems for MICE:

- Circular dependence can occur (e.g., $X_1 \Leftrightarrow X_2$).
- Multicollinearity problem.
- Imputation can generate impossible combinations (e.g., pregnant father).
- Imputation can destroy deterministic relations in the data (e.g., sum scores).

Considerations for Imputation

- Imputation should be performed prior to any pre-processing methods.
- By the rule of thumb, less than 20% of missing values for each variable would be good for imputation.
- Using Y to impute X's would result in circular analysis $\Rightarrow$ Importance of X_j with imputed values can be exaggerated.
- Importance of X_j with relatively many imputed values:
  - Sensitivity analysis: Comparison of models with and without obs. with imputed values for the X_j.

Considerations for Multiple Imputation

- Statistical imputation methods are mainly concerned with the impact on statistical inference ⇒ Good enough quality to support distributions of statistics and hypothesis tests. ⇒ Multiple imputation.
- Impact of imputation for predictive models:
  - The multiple imputation may not have relevance for predictive models without distributional assumptions.
  - For predictive models required expensive computing, the multiple imputation could be burdensome. Models with resampling should have imputation within resampling process.
  - Since predictive models mainly focus on accurate prediction of new unseen objects, imputed values should be as close as possible to their true values.
  - Multiple imputation does not keep the imputation generator after imputation. ⇒ Difficult to apply the imputation to new unseen obs. with missing values.