Chapter 3. Preparing the Input Variables

3.1 Missing Values

Does Pr(missing) Depend on the Data?

- No
  - MCAR (Missing Completely at Random)
    - ✓ a particularly easy mechanism to manage but is unrealistic in most predictive modeling applications
- Yes
  - depend on the unobserved value
    - ✓ ex. Credit applicants with fewer years at their current job might be less inclined to provide this information.
  - depend on values of other unobserved values
    - ✓ ex. Transient customers might have missing values on a number of variables.
  - depend on observed values of other input variables (including the target)
    - ✓ ex. Customers with longer tenures might be less likely to have certain historic transactional data.

➔ A fundamental concern for predictive modeling is that the missingness is related to the target. The more transient customers might be the best prospects for a new offer.

Complete Case Analysis

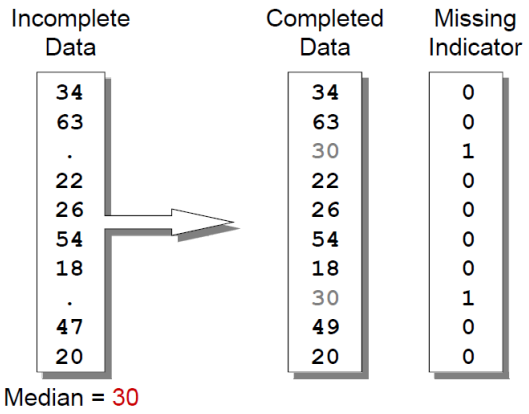
In *complete-case analysis,* only those cases without any missing values are used in the analysis. Complete-case analysis has some moderately attractive theoretical properties even when the missingness depends on observed values of other inputs. However, complete-case analysis has serious practical shortcomings with regard to predictive modeling. Even a smattering of missing values can cause an enormous loss of data in high dimensions.

Missing Value Imputation

Because of the drawbacks of complete-case analysis, some type of missing value imputation is necessary. Imputation means filling in the missing values with some reasonable value. Many methods have been developed for imputing missing values. The principal consideration for most methods is getting valid statistical inference on the imputed data, not generalization.

Imputation + Indicators

One reasonable strategy for handling missing values in predictive modeling is to do the following steps.



1. Create missing indicators and treat them as new input variables in the analysis.
2. Use median imputation for numeric inputs. Fill the missing values of x_j with the median of the complete cases for that variable.
3. Create a new level representing missing (unknown) for categorical inputs.

➔ There is a large amount of statistical literature considering different missing value imputation methods, including discussions of the demerits of mean and median imputation and missing indicators. Unfortunately, there is very little advice when the functional form of the model is not assumed to be perfectly specified, when the goal is to get good predictions that can be practically applied to new cases, when p-values and hypothesis tests are of secondary importance, and when the missingness might be highly pathological, in other words, depending on lurking predictors.

Cluster Imputation (군집 평균 대체)

Mean imputation uses the unconditional mean of the variable. An attractive extension would be to use the mean conditional on the other input, referred to as regression imputation. Regression imputation would usually give better estimates of the missing values. However, a complication of regression imputation is that the other inputs might themselves have missing values. Consequently, the *k* imputation regressions also need to accommodate missing values.

Cluster-mean imputation is a somewhat more practical alternative:

1. cluster the cases into relatively homogenous subgroups
2. mean-imputation within each group
3. for new cases with multiple missing values, use the cluster mean that is closets in all the non-missing dimensions.

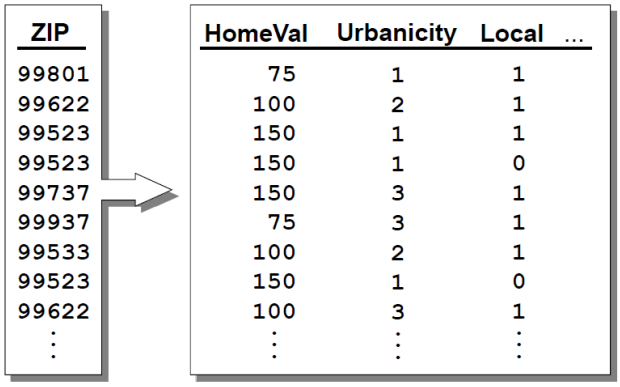
A simpler but sometimes useful alternative is to define a priori segments (for example, high, middle, low, and unknown income), and then do mean or median imputation within each segment.

3.2 Categorical Inputs

Dummy Variables (가변수)

A dummy variable (aka. indicator variable) is a numeric variable that represents categorical data, such as gender, race, political affiliation, etc. To represent a categorical variable that can assume *k* different values, a researcher would need to define *k*-1 dummy variables.

Smarter Variables



ZIP
99801
99622
99523
99523
99737
99937
99533
99523
99622
⋮

HomeVal	Urbanicity	Local	...
75	1	1	
100	2	1	
150	1	1	
150	1	0	
150	3	1	
75	3	1	
100	2	1	
150	1	0	
100	3	1	
⋮	⋮	⋮	

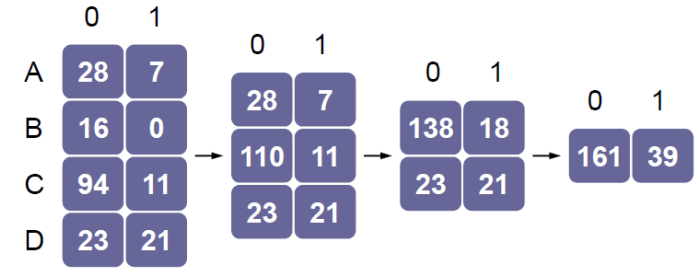
Expanding categorical inputs into dummy variables can greatly increase the dimension of the input space. A smarter method is to use subject-matter information to create new inputs that represent relevant sources of variation. For example, geographic areas are often mapped to several relevant demographic variables.

Quasi-Complete Separation

Quasi-complete separation occurs when a level of the categorical input has a target event rate of 0 or 100%. When quasi-complete separation occurs, one of the logits will be infinite. This complicates model interpretation. Furthermore, it might lead to incorrect decisions regarding variable selection.

The most common cause of quasi-complete separation in predictive modeling is categorical inputs with rare categories. The best remedy for sparseness is collapsing levels of the categorical variable (범주 병합).

Clustering Levels



Merged:	B & C	A & BC	ABC & D
χ^2 =	31.7	30.7	28.6	0
	100%	97%	90%	0%

Ideally, subject-matter considerations should be used to collapse levels of categorical inputs, but this is not always practical in predictive modeling. A simple data-driven method was developed by Greenacre, where the levels are hierarchically clustered based on the reduction in the chi-squared test of association between the categorical variable and the target.

At each step, the two levels that give the least reduction in the chi-squared statistic are merged. The process is continued until the reduction in chi-squared drops below some threshold.

This method will quickly throw rare categories in with other categories that have similar marginal response rates. While this method is simple and effective, there is a potential loss of information because only univariate associations are considered.

3.3 Variable Clustering

Redundancy

Including redundant inputs can degrade the analysis by

- destabilizing the parameter estimates
- increasing the risk of overfitting
- confounding interpretation
- increasing computation time
- increasing scoring effort
- increasing the cost of data collection and augmentation

Redundancy is an *unsupervised* concept that does not involve the target variable. On the other hand, the relevancy of a variable takes into account the relationship between an input variable and the target variable. In high-dimensional data sets, identifying irrelevant inputs is more difficult than identifying redundant inputs. A good strategy is to first reduce redundancy and then tackle irrelevancy in a lower dimension space.

Variable Clustering

Variable clustering finds groups of variables that are as correlated as possible among themselves and as uncorrelated as possible with variables in other clusters.

Principal Components (주성분 분석)

$$PC_{(1)} = W_{(1)1}X_1 + W_{(1)2}X_2 + \ldots + W_{(1)p}X_p$$
$$PC_{(2)} = W_{(2)1}X_1 + W_{(2)2}X_2 + \ldots + W_{(2)p}X_p$$
$$\vdots$$
$$PC_{(p)} = W_{(p)1}X_1 + W_{(p)2}X_2 + \ldots + W_{(p)p}X_p$$

where the weights have been chosen to maximize the quantity

$$\frac{\text{Variance of PC}}{\text{Total Variance}}$$

and the correlation $corr(PC_{(i)}, PC_{(j)})=0$ for each i not equal to j.

Principal components are weighted linear combinations of the predictor variables where the weights are chosen to account for the largest amount of variation in the data. The principal components are numbered according to how much variation in the data is accounted for and each principal component accounts for a unique portion of the variation in the data. In other words, they are not correlated.

각 PC는 변동을 설명하는 정도에 따라 순차적으로 넘버링이 되고 각 PC는 독립!

The *eigenvalues* are the variances of the PCs; they sum to the number of variables. Each PC explains a decreasing amount of the total variability.

Principal Component Coefficients

The coefficients of the PCs (eigenvectors) are usually nonzero for all the original variables. Thus, even if only a few PCs were used, all the inputs would still have to be retained in the analysis.

The VARCLUS Procedure

```
PROC VARCLUS DATA=SAS-data-set <options>;
    VAR variables;
RUN;
```

Selected PROC VARCLUS statement options:

MAXEIGEN=*n* specifies the largest permissible value of the second eigenvalue in each cluster. The default is 1 (using the correlation matrix).

SHORT suppresses printing of the cluster structure, scoring coefficient, and intercluster correlation matrices.

Selected VARCLUS procedure statement:

VARCLUS specifies the variables to be clustered. If you do not specify the VAR statement, all numeric variables not listed in other statements are processed.

Cluster Representatives

$$1 - \mathrm{R}^2\ ratio = \frac{1 - R^2_{own\ cluster}}{1 - R^2_{next\ cluster}}$$

smaller is better!!

As with principal components analysis, dimension reduction could be achieved by replacing the original variables with the cluster scores (components). An ideal representative would have high correlation with its own cluster and a low correlation with the other clusters. Consequently, variables with the lowest $1 - \mathrm{R}^2\ ratio$ in each cluster would be good representatives.

3.4 Variable Screening

Univariate Screening

It is tempting to use univariate associations to detect irrelevant input variables, where each input variable is screened individually versus the target variable. Only the most important inputs are retained in the analysis. This method does not account for partial associations among the inputs. Input could be erroneously omitted or erroneously included.

Because some of the variable selection techniques use the full model, eliminating clearly irrelevant variables will stabilize the full model and might improve the variable selection technique without much risk of eliminating important input variables. However, you should keep in mind that univariate screening can give misleading results when there are partial associations.

Empirical Logits

In regression analysis, it is standard practice to examine scatter plots of the target versus each input variable. However, when the target is binary, these plots are not very enlightening. A useful plot to detect nonlinear relationships is al plot of the empirical logits.

$$\ln\left(\frac{\mathrm{m}_i + \frac{\sqrt{M_i}}{2}}{M_i - m_i + \frac{\sqrt{M_i}}{2}}\right)$$

where $\mathrm{m}_i = number\ of\ events$

$\mathrm{M}_i = number\ of\ cases$

Univariate plots of binary data need to be smoothed in order to better reveal the relationship between a continuous input variable and the target. A simple, scalable, and robust smoothing method is to plot empirical logits for quantiles of the input variables.

Remedies

1. Hand-Crafted New Input Variables
2. Polynomial Models
3. Flexible Multivariate Function Estimators
4. Do Nothing

3.5 Subset Selection

All Subsets

Variable selection methods in regression are concerned with finding subsets of the inputs that are jointly important in predicting the target. The most thorough search would consider all possible subsets. This can be prohibitively expensive when the number of inputs, k, is large, as there are 2^k possible subsets to consider.

Stepwise Selection

Stepwise variable selection is an often criticized and yet heavily used subset selection method. Stepwise selection searches the input models and selects the best. The model is incrementally built in this fashion until no improvement is made. There is also a backward portion of the algorithm where at each step, the variables in the current model can be removed if they have become unimportant.

Stepwise selection was devised to give a computationally efficient alternative to examining all subsets. It is not guaranteed to find the best subset and it can be shown to perform badly in many situations.

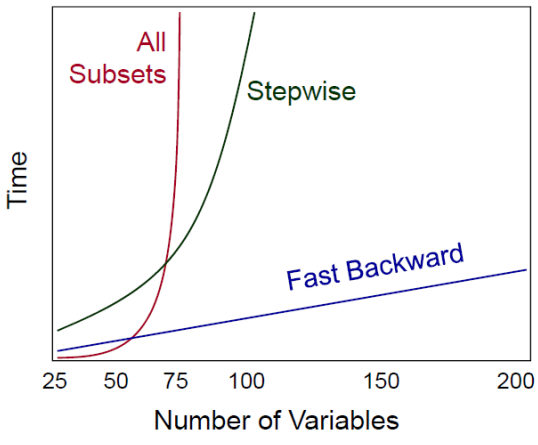
Backward Elimination

Backward variable selection starts with all the candidate variables in the model simultaneously. At each step, the least important input variable is removed. Backward elimination is less inclined to exclude important inputs or include spurious inputs than forward (stepwise) methods. However, it is considered more computationally expensive than stepwise because more steps are usually required and they involve larger models.

Scalability in PROC LOGISTIC

The conventional wisdom regarding computation time is that

stepwise < backwards < all subsets.



However, logistic regression gives a different story. For up to ≈ 60 inputs, the results are reversed **all subsets < backwards < stepwise**.

For any number of inputs, backward elimination is more efficient than stepwise. Logistic regression requires an iterative optimization algorithm. Each step in the stepwise algorithm requires iterative optimization. All-subsets selection is the fastest until the number of possible combinations become unmanageable, at which point the performance acutely deteriorates. If redundant inputs are eliminated first, then all-subsets selection can be a practical method for predictive modeling.

3.6 Chapter Summary

General form of the **STDIZE** procedure:

```
PROC STDIZE DATA=SAS-data-set <options>;
    VAR variables;
RUN;
```

General form of the **CLUSTER** procedure:

```
PROC CLUSTER DATA=SAS-data-set <options>;
    FREQ variable;
    VAR variable;
    ID variable;
RUN;
```

General form of the **VARCLUS** procedure:

```
PROC VARCLUS DATA=SAS-data-set <options>;
    VAR variables;
RUN;
```