3.5 Chapter Summary

Preparing the data for predictive modeling can be laborious. First, missing values need to be replaced with reasonable values. Missing indicator variables are also needed if missingness is related to the target. If there are nominal input variables with numerous levels, the levels should be collapsed to reduce the likelihood of quasi-complete separation and to reduce the redundancy among the levels. Furthermore, if there are numerous input variables, variable clustering should be performed to reduce the redundancy among the variables. Additionally, there are several selection methods in the LOGISTIC procedure to select a subset of variables.

To assist in identifying nonlinear associations, the Hoeffding's D statistic can be used. A variable with a low rank in the Spearman correlation statistic but with a high rank in the Hoeffding's D statistic may indicate that the association with the target is nonlinear.

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;
```