Read About It: Partitioning a Data Set Using PROC GLMSELECT

If you start with a data set that's not yet partitioned, PROC GLMSELECT can partition the data for you. You can request two partitions (training and validation) or three partitions (training, validation and testing). You specify the proportion to use for the validation and test data cases, and you can specify a seed for the partitioning algorithm.

Partitioning a Data Set Using PROC GLMSELECT

```
PROC GLMSELECT DATA=training-data-set <SEED=number>;
      MODEL targets=inputs < / options>;
      PARTITION FRACTION(<TEST=fraction> <VALIDATE=fraction>) ;
RUN;
```

In the PROC GLMSELECT statement, the DATA= option specifies the input or training data set. You'll use the PARTITION statement to specify how the cases in the input data set are partitioned into holdout samples for model validation, and if desired, testing. The MODEL statement is the same as before.

The PARTITION statement specifies how observations in the input data set are logically partitioned into disjointed subsets for model training, validation, and testing. The FRACTION option specifies the fraction (that is, the proportion) of cases in the input data set that are randomly assigned to a testing role and a validation role. The sum of the specified fractions must be less than 1 and the remaining fraction of the cases in the input data set are assigned to the training role. For example, the statement below requests two partitions (training and validation), and one quarter, or 25%, of the observations are written to the validation data set. The remaining three quarters, or 75%, are written to the training data set.

```
PARTITION FRACTION(VALIDATE=.25);
```

The PARTITION statement uses a pseudo-random number generator. To begin the random selection process, it needs a starting "seed," which must be an integer. If you want to reproduce your results in the future, specify an integer greater than zero in the SEED= option. Then, whenever you run the PROC GLMSELECT step and use the same seed value, the selection process is replicated and the same results are generated. If the SEED= value is invalid or omitted, the seed is automatically generated from the computer's clock. In most situations, it's recommended that you use the SEED= option and specify an integer greater than zero.

Partitioning a Data Set Using the Predictive Regression Models Task

You can use the Predictive Regression Models task to partition a data set into two or three partitions. If you want two partitions (training and validation), you must specify a sample proportion for the validation cases. This required value, which is a number between 0 and 1, represents the fraction or proportion of observations to be written to the validation partition. The remaining observations are written to the training partition.

If you also want a test partition, then you indicate that in the task, and specify a sample proportion for the testing partition. This value (a number between 0 and 1) represents the fraction or proportion of cases to be written to the testing partition. If you request both validation and testing partitions, then the sum of the specified fractions must be less than one. The remaining observations are written to the training partition.

You can use the random seed option to specify a starting seed for the pseudo-random number generator. If you specify an integer that's greater than zero, you can reproduce the results in the future. If you omit this option, a random seed will be generated, and the results will be different each time you submit the code.

Type and save a note for this page.

Submit

Close