

検証列の作成(ホールドアウト法)

モデルを作成するためのサブセット(学習セット)とモデルの予測性能を評価するためのサブセット(検証セット)に、データを分割します。複数のモデルが作成されたときに、検証データで最も性能の高いモデルがしばしば選択されます。また、新しいデータ上でモデルの予測性能を評価するために、3つ目のサブセット(テストセット)を用いることもあります。テストセットはモデルの作成にも選択にも用いられていないため、モデルの予測性能を評価する上で、この方法がより正確な方法であると考えられています。

データに過適合する傾向のあるモデルを作成する際に、検証列を使用すると特に有用です。JMP のいくつかのモデル作成用のプラットフォームでは、モデルをあてはめる時に検証データの割合を指定するオプションがあるため、検証列の作成は不要です。

JMP Pro での検証列の作成(学習、検証、テスト)

1. データテーブルで、**分析 > 予測モデル > 検証列の作成**を選択します。
2. 層別の列、グループの列、カットポイントの列を指定して、分割手法を選択できます。単純無作為抽出を用いて検証列を作成する場合は、**OK** をクリックします。
3. 表示されるウィンドウで、データを学習セット、検証セット、テストセットへ割り当て方を決めるために値(度数もしくは割合)を入力します。同じ無作為割付を再現する場合は、乱数シード値を指定します。

新しい列が作成され、指定した割合(もしくは度数)で0、1、2 という値が入力されます。



検証列の作成

単純無作為抽出による検証列

データテーブルの各行を、学習・検証・テストの各セットにランダムに分ける。学習セットは、モデルの推定に使う。複数の候補モデルの予測能力を比較するために用いる。検証セットは、選択されたモデルの予測能力を独立して評価するために用いる。テストセットの指定は任意である。

割合、または、割合の相対的な大きさを指定する。

	調整後の割合	行数
学習セット	0.6	0.6 3576
検証セット	0.3	0.3 1788
テストセット	0.1	0.1 596
除外されている行		0
全体の行数		5960

オプション

新しい列の名前: 検証

検証列の種類: 固定値

乱数シード値:

実行

キャンセル ヘルプ

- 学習セットの 3,576 例(60%)は、モデルを構築(訓練)するのに用いられます。
- 検証セットの 1,788 例(30%)は、最良のモデルを検証して選択するのに用いられます。
- テストセットの 596 例(10%)は、新しいデータを用いて、選択したモデルの性能をテストするのに用いられます。

JMP での検証列の作成

1. データテーブルで、**列メニューから列の新規作成**を選択します。
2. **列の新規作成**ウィンドウで、**列名を検証**に変更します。
3. **データの初期化**の選択から**乱数**を選択します。
4. **指示乱数**を選択します。希望する割合を入力します。この例では、50%の0(学習)、30%の1(検証)、20%の2(テスト)を選びます。
5. 0、1、2 に対して学習、検証、テストのラベルを割り当てて表示する場合、**列プロパティ**以下の**値ラベル**を選択し、それぞれの値にラベルを割り当てて**追加**をクリックします。
6. **適用**をクリックして、データテーブル上の新しい列を確認します(列が指定したように作成されたかを確認します)。その後 **OK** をクリックします。



データの初期化

乱数

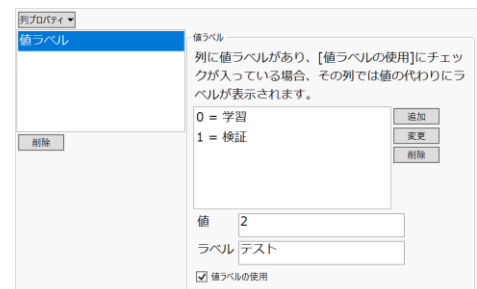
整数乱数 値 割合

一様乱数 0 0.5

正規乱数

指示乱数 1 0.3

2 0.2



列プロパティ

値ラベル

列に値ラベルがあり、[値ラベルの使用]にチェックが入っている場合、その列では値の代わりにラベルが表示されます。

0 = 学習

1 = 検証

値 2

ラベル テスト

値ラベルの使用

追加

変更

削除

削除

注意: 検証列の作成の追加情報に関しては、**基本的な回帰モデル、予測モデルおよび発展的なモデル(ヘルプ > JMP ドキュメンテーションライブラリ以下)**をご覧ください。もしくは、JMP のヘルプを「検証」を検索してご確認ください。