mwp-444

Lasso intro - Lasso の機能紹介　【 評価版 】

本 whitepaper では lasso の機能概要について紹介します。

1. Lasso とは？

Lasso というのは元々“least absolute shrinkage and selection operator”の略語だったわけですが、最近では単独のキーワードとして用いられるようになっています。

Lasso とはモデル中に現れる共変量の中から選択を行い、それらをフィットさせる手法のことを言います。Stata の `lasso` コマンドは線形/ロジット/プロビット/ポアソンモデルのフィットを行うことができます。今、従属変数 `y` を共変量 `x1, x2, ..., x`p によって説明しようとする線形モデルについて考えることにしましょう。通常であれば

```
. regress y x1 x2 ... xp
```

のようにコマンド入力することになります。

今、どの変数（共変量）をモデルに含めたら良いか確信が持てないとします。ただし変数の中には間違いなくモデルに含めるべきものがいくつか存在し、その数は観測データ数 N に比べて小さいことも仮定します。その場合、

```
. lasso linear y x1 x2 ... xp
```

という形でモデルのフィットを行うことができます。

共変量の数としては数百、あるいは数千といったものが許容されます。観測データの総数よりも多くの共変量があっても構いません。その中から適切なものを lasso が選び出すことになります。

Lasso の用途としては次の 3 つがあります。

1. 予測 - 多数の回帰変数を前提とした上でアウトカム変数の値を予測する。
2. モデル選択 - アウトカム変数をうまく予測できる変数の集合を選択する。
   真のモデル中に含まれる変数を選択しようとするものではありません。また係数値についての科学的な解釈を導こうとするものでもありません。1 つのデータセット中のアウトカム変数と良い相関を示す変数群を選択すると共に、それらが他のデータセット中のアウトカム変数をもうまく予測できるか否かを検証することを目的とするものです。
3. 推論 - フィットされたモデルの係数値の意味を解釈するために推論を行う。
   真のモデル中における変数の効果の推定、標準誤差、信頼区間、p 値、等の推定が含まれます。

2. **予測のための** Lasso

評価版では割愛しています。

3. **モデル選択のための** Lasso

評価版では割愛しています。

4. **推論のための** Lasso

評価版では割愛しています。

■