第9回ベイズ統計学・機械学習研究会

Kyohei ITO

2020/11/9

ロジスティック回帰

説明変数から確率を計算し予測する一般化線形モデルの一種です。「回帰」と名付けられているが分類に使います。今回は０か１で予測する二項ロジスティック回帰を扱います。3つ以上に分類するときは多項ロジスティック回帰を用います。

一般化線形モデルとは

一般線形モデル(回帰分析で使ったやつ)では目的変数が正規分布に従うことを前提としていましたが、一般化線形モデルでは目的変数が正規分布に従わなくても適用でき、さらに質的変数であっても使用できます。

$$g(y) = \beta_0 x_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

$g(\cdot)$関数はリンク関数と言います。

リンク関数

リンク関数は被説明変数を予測可能な形に変形させるものです。被説明変数の分布に対して一意に決まります。

分布族(family)	リンク関数 g(μ)	y_i の範囲
正規(gaussian)	μ	(-∞,+∞)
二項(binomial)	log(μ/(1-μ))	$\frac{0,1,2,\ldots,n_i}{n_i}$
ポアソン(poisson)	log(μ)	0, 1, 2, …
ガンマ(Gamma)	1/μ	(0,+∞)
逆正規(Inverse.gaussian)	$1/\mu^2$	(0,+∞)

今回は二値分類なので二項分布に対応するリンク関数を適応します。

多重共線性

重回帰分析と同様、多重共線性への注意が必要です。相関係数行列などで確認しましょう。

Rのコード

ロジスティック回帰

```
glm(y~.,data,family=binomial)
```

モデルのあてはめ

type = “response”とすることに注意してください。

```
pred_1 <- predict(model1,newdata = df_test,type = "response")
```

欠損値の除去

今回のデータは欠損値を含みます。欠損値の処理方法は色々ありますが、今回は欠損値を含む列を削除する方法を取ります。

```
drop_na(Data)
```

閾値（しきいち）の決定

ロジスティック回帰は1の値を取る確率を求めてきます。何%以上なら1、それ以下なら0とする基準（閾値）を自分で決めなければいけません

```
Data %>%
  mutate(Predicted_value = if_else(column > 0.5, 1, 0))
```

モデルの評価

予測値と実測値のクロス集計表を作成することで正解率を求めることができます。

```
table("予測値"=col,"実測値"=col)
```

今回使用するデータセット

URL:https://www.kaggle.com/dileep070/heart-disease-prediction-using-logistic-regression?select=framingham.csv (https://www.kaggle.com/dileep070/heart-disease-prediction-using-logistic-regression?select=framingham.csv)

心臓病の予測

•性別：男性または女性（名目）

•年齢：患者の年齢;（継続的-記録された年齢は整数に切り捨てられていますが、年齢の概念は継続的です）

行動

•現在の喫煙者：患者かどうかは現在喫煙者です（名目）

•1日あたりのタバコの数：1日に平均して喫煙したタバコの数（1本は任意の数のタバコ、半分のタバコでも持つことができるため、継続的と見なすことができます。）

医療（履歴）

•BPMeds：患者が血圧の薬を服用していたかどうか（名目）

•流行のストローク：患者が以前に脳卒中をしたかどうか（名目）

•流行のHyp：患者が高血圧であったかどうか（名目）

•糖尿病：患者が糖尿病であるかどうか（名目）

医療（現在）

•Tot Chol：総コレステロールレベル（連続）

•Sys BP：収縮期血圧（連続）

•Dia BP：拡張期血圧（連続）

•BMI：体重指数（連続）

•心拍数：心拍数（連続-医学研究では、心拍数などの変数は実際には離散的ですが、可能な値が多数あるため連続と見なされます。）

•グルコース：グルコースレベル（連続）

予測変数

•冠状動脈性心拍数CHDの10年リスク（バイナリ：「1」、「はい」を意味し、「0」は「いいえ」を意味します）

TenYearCHDをyとして予測してください。