データ解析のための統計学入門

第4章
GLMのモデル選択
—AICとモデルの予測の良さ—

山木翔馬

モデル選択

- 複数のモデルの中から, 何らかの意味で「良い」モデルを選ぶことをモデル選択（model selection）という
- 観測データへのあてはまりの良さを選択基準にする場合
  - ➔ 最大対数尤度
  - ➔ モデルを複雑にすればあてはまりは良くなる
  - ➔「良いモデル」とは言えない
- AIC
  - ➔「良い予測をするモデルが良いモデル」という考えにもとづいた選択基準

逸脱度（deviance）

- 「あてはまりの悪さ」を表現する指標
- 対数尤度を $\log L$, 最大対数尤度を$\log L^*$ としたとき

$$D = -2 \log L^*$$

名前	定義
逸脱度（D）	$-2 \log L^*$
最小の逸脱度	フルモデルをあてはめたときのD
残差逸脱度	D - 最小のD
最大の逸脱度	Null モデルをあてはめたときのD
Null 逸脱度	最大のD - 最小のD

glm( )の出力

- 第3章の3.4.1項で使った, 平均種子数 λ_i が植物の体サイズ x_iだけに依存するモデル $\lambda_i = exp(\beta_1 + \beta_2 x_i)$ を「xモデル」という

  xモデルをglm( )にあてはめた出力

Null Deviance: 89.51

Residual Deviance: 84.99 AIC: 474.8

- 残差逸脱度（residual deviance）＝

  D -（ポアソン分布モデルで可能な最小逸脱度）
- ポアソン分布モデルで可能な最小の逸脱度 ＝

  フルモデルの逸脱度

フルモデル

- データ数と同数のパラメータを使って「あてはめた」モデル
- ポアソン回帰で可能な他のどのモデルを使った場合よりも, 対数尤度は大きくなる
- 3章のデータのフルモデルの対数尤度は

  > sum(log(dpois(d$y, lambda = d$y)))

  [1] -192.8898
- 逸脱度 D = 385.8 がこのデータのもとで, ポアソン回帰で可能な最小逸脱度
- xモデルの最大対数尤度 -235.4（3章より）,　逸脱度 470.8

  残差逸脱度 = 470.8 – 385.8 = 85.0

逸脱度の最大値

- 「もっともあてはまりの悪いモデル」のとき, 逸脱度が最大
- 例題の場合, もっともパラメータ数の少ないモデル
  - 平均種子数が $\lambda_i = exp(\beta_1)$ と指定されているだけのモデル（パラメータ数 k = 1）
  - Rでは null model とよばれる

```
> fit.null <- glm(formula = y ~ 1, family = poisson, data = d)

Null Deviance:      89.51

Residual Deviance: 89.51    AIC: 477.3
```

- このデータを使ったポアソン回帰での残差逸脱度の最大値は89.5になる

- 最大対数尤度は -237.6

>logLik(fit.null)

'log Lik. ' -237.6432 (df=1)

- 逸脱度は 475.3
- この逸脱度と最小D（385.8）の差が89.5ぐらいになる

- パラメータ数kを増やせば残差逸脱度は小さくなり、あてはまりが良くなる

AIC（Akaike's information criterion）

- 予測の良さを重要視するモデル選択基準
- 最尤推定したパラメータ数がkであるとき

  AIC = -2 {（最大対数尤度）-（最尤推定したパラメータ数）}

  $= -2 (\log L^* - k)$

  $= D + 2k$
- AICが一番小さいモデルが良いモデルとなる

平均対数尤度

- 最大対数尤度は真の統計モデル（観測データを生成したモデル）へのあてはまりの良さではなく, パラメータ推定に使った観測データへのあてはまりの良さ
- 推定したモデルが真のモデルにどれくらい近いか調べるには, 推定に使った観測データとは別の観測データによる検証が必要
- 推定に使った観測データとは別の観測データに対する対数尤度の平均が平均対数尤度 E(logL) となる

最大対数尤度のバイアス補正

- 最大対数尤度 logL*と平均対数尤度E(logL)の差

  b = logL* - E(logL) をバイアスという
- E(logL) = logL* - b とすれば, 平均的なbと最大対数尤度 logL*がわかれば平均対数尤度の推定量が得られる

  →バイアス補正
- 解析的に平均対数尤度の推定量は logL* - k

ネストしているGLM間のAIC比較

- $log\lambda_i = \beta_1$ …… 一定モデル（k = 1）
- $log\lambda_i = \beta_1 + \beta_2 x_i$ …… xモデル（k = 2）

※説明変数xiは応答変数とは全く無関係

- AICを基準に考えると，xモデルは一定モデルよりパラメータが多いため，予測の良さは悪化する
- 最大対数尤度を基準に考えると，β_2 をうまく選んでしまえば一定モデルよりxモデルのほうが「あてはまり」は良くなる

あてはまりの良さの指標の改善だけをめざしたモデルの複雑化は危険