応用統計学 2017 第 11 回 一般化線形モデル

2017 年 12 月 20 日 (水)
清 智也 sei@mist.i.u-tokyo.ac.jp
http://ur0.pw/yTzt

- 一般化線形モデル，線形予測子，リンク関数，ロジスティック回帰，ポアソン回帰[1]。
- 参考文献[2]
  - 汪金芳「一般化線形モデル」朝倉書店
  - 久保拓弥「データ解析のための統計モデリング入門」岩波書店

表 1: 一般化線形モデル（主なもの）をまとめた一枚の表。

		モデル	説明変数 $\boldsymbol{x}$ の種類による区別				
			離散 1 変数 2 水準	離散 1 変数 多水準	離散多変数	連続 1 変数	連続多変数
目的変数 y	連続値	**正規線形モデル**	平均値の差を t 検定	平均値の差を 分散分析	交互作用を 分散分析	単回帰 分析	重回帰 分析
	あり／なしなどの二値	**ロジスティック回帰**	対数線形モデル			ロジスティック 回帰	
	度数	**ポアソン回帰**				ポアソン回帰	

西内 啓「統計学が最強の学問である」p.170 の表を一部改変。

演習問題

問題 11-1. ロジスティック回帰モデルは，$\{0, 1\}$ 値の目的変数 Y_t に対して

$$Y_t \sim \text{Bernoulli}(\mu_t), \quad \log\left(\frac{\mu_t}{1-\mu_t}\right) = \boldsymbol{\beta}'\boldsymbol{x}_{(t)}, \quad t = 1, \ldots, n,$$

と定義される。ただし $\boldsymbol{x}_{(t)}$ は説明変数であり，$\boldsymbol{\beta}$ はパラメータである[3]。尤度関数を求めよ。

問題 11-2. ポアソン回帰モデルは，非負整数値の目的変数 Y_t に対して

$$Y_t \sim \text{Poisson}(\mu_t), \quad \log \mu_t = \boldsymbol{\beta}'\boldsymbol{x}_{(t)}, \quad t = 1, \ldots, n,$$

と定義される。ただし $\boldsymbol{x}_{(t)}$ は説明変数であり，$\boldsymbol{\beta}$ はパラメータである。尤度関数を求めよ。

[1] generalized linear model, linear predictor, link function, logistic regression, Poisson regression.
[2] 対数線形モデルについては宮川「統計技法」にも説明がある。
[3] 重回帰モデルのときと同じく，切片を陽に書いて $\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_{(t)}$ とすることも多い。本稿では表記の簡単のため $\beta_0 = 0$ とし，代わりに $\boldsymbol{x}_{(t)}$ の一つの成分が定数 1 であると考えることにする。以下同様。

問題 11-3. （カノニカルな）一般化線形モデルは，目的変数 Y_t と説明変数 $\boldsymbol{x}_{(t)}$ に対して

$$f(y_t) = a(y_t, \phi) \exp\left(\frac{\theta_t y_t - \psi(\theta_t)}{\phi} \right), \tag{1}$$

$$\theta_t = \boldsymbol{\beta}' \boldsymbol{x}_{(t)}, \tag{2}$$

と定義される。ただし，$a(y, \phi)$ と $\psi(\theta)$ は関数であり，パラメータは $\boldsymbol{\beta}$, $\phi > 0$ である[4]。式 (1) は ϕ を固定すると指数型分布族になっている。また式 (2) は線形予測子と呼ばれる。

(i) Y_t の平均と分散は $\mu_t := \mathrm{E}[Y_t] = \psi'(\theta_t)$，$\mathrm{Var}[Y_t] = \phi\psi''(\theta_t)$ と表されることを示せ。線形予測子を平均で表した関数 $\theta_t = (\psi')^{-1}(\mu_t)$ のことをリンク関数という。リンク関数を用いると式 (2) は $(\psi')^{-1}(\mu_t) = \boldsymbol{\beta}' \boldsymbol{x}_{(t)}$ と書ける[5]。

(ii) 正規線形モデル，ロジスティック回帰モデル，ポアソン回帰モデルはいずれも一般化線形モデルであることを示せ。 ただしそれぞれ $\phi = \sigma^2$, $\phi = 1$, $\phi = 1$ とする。

問題 11-4. ポアソン回帰モデルにおいて，総度数 $\sum_{t=1}^{n} Y_t = \nu$ が固定されたもとでの条件付き分布を求めると，どのような統計モデルが得られるか。

問題 11-5. 第 3 回，第 4 回の宿題で扱った「世界の上位 50 大学」の例について，アメリカの大学とそれ以外を判別する判別関数を，ロジスティック回帰モデルに基づいて求めよ。

宿題 11

問題 11-6. 講義ページにあるデータファイル `FIFA.csv` は，web サイト

FIFA Ranking.net (http://fifaranking.net/nations/jpn/)

の情報をもとに，サッカー日本代表の 2014 年から 2017 年の戦績を表にしたものである。各変量の意味は次の通りである：

date	opponent	goal1	goal2	stadium	rank1	rank2
日付	対戦相手	得点	失点	ホーム/アウェイ	日本の順位	相手国の順位

いま，ホーム/アウェイと両チーム順位から得点を予測するモデルを作りたいものとする。次の R プログラムを実行し，その出力結果の意味を説明せよ。

```
> X = read.csv("FIFA.csv")
> glm.1 = glm(goal1 ~ stadium + rank1 + rank2, family=poisson, data=X)
> summary(glm.1)
```

[4] ϕ は定数の場合と未知パラメータの場合がある。また，関数 $a(y, \phi), \psi(\theta)$ は自由に選べるわけではなく，$a(y, 1)$ を定めれば高々一つに定まることが知られている。

[5] 式 (2) の代わりに何らかの関数 g を用いて $g(\mu_t) = \boldsymbol{\beta}' \boldsymbol{x}_{(t)}$ と仮定するモデルを，カノニカルでない一般化線形モデルという。このときの g もリンク関数と呼ぶ。