データの当てはまり具合を評価する

黒沢 健 (東京理科大学 応用数学科)

1　数理モデル

世の中には数理現象が沢山ある. 例えば

- 「歩数と移動距離の関係」
- 「動画を見た時の電池の残時間」

等が挙げられる. これらの数理現象を実際に計測したデータを用いて解析し, 統計モデルを作る. そしてその統計モデルを使い,「歩数から移動距離の推定」や「動画を見た時の電池の残時間の推定」を行う.

しかしながら, 観測には誤差が発生する場合がある. どういった基準で最も良いモデルを作り, どうやって良いものを選ぶのか, また統計モデルがもっともらしいものかをチェックする必要がある. そのもっともらしさを判断する為の基準であるモデル評価尺度について考える.

2　線形モデルの選択

10 人の学生による統計学と経済学の期末テストのデータセット (統計学, 経済学) $= (62, 66), (45, 48), \ldots, (81, 75)$ が得られたとしよう.

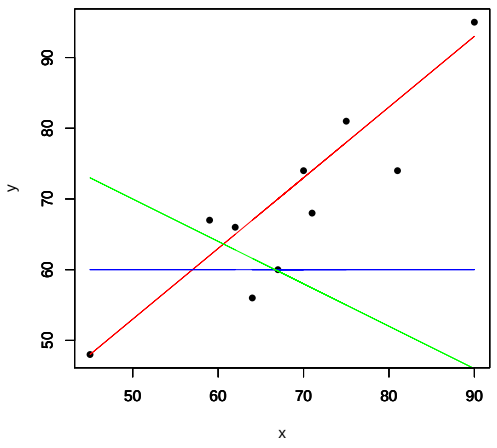
得られたデータを元に行う事

統計学と経済学の期末テストに関する 10 人のデータを用いて, 統計学の点数から経済学の点数を推定する良い統計モデルを作る.

x を統計学の点数, y を経済学の点数とした時に

$$y = a + bx$$

という統計モデルについて考える.



統計モデル赤, 青, 緑のうちどれが良いでしょうか?

- $y = x + 3$
- $y = 60$
- $y = -0.6x + 100$

- どうやって決めましたか?
- 何をもって近いと判断しましたか?

観測データ数 $n = 10$ とし, $\hat{y}_i$ を統計モデルの推定値 $a + bx_i$ $(1 \le i \le n)$ とし, $\bar{y}$ を 10 人の経済学の点数の平均値 としたとき,

$$\hat{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1}$$

を計算する. この $\hat{R}^2$ を赤, 青, 緑の統計モデルに対して計算し, 一番値が大きい統計モデルを選択する.

- 赤: $y = x + 3$ に関して $\hat{R}^2 = 0.87$.
- 青: $y = 60$ に関して $\hat{R}^2 = 0$.
- 緑: $y = -0.6x + 100$ に関して $\hat{R}^2 = 0.31$.

よって $\hat{R}^2$ が一番大きい赤が良い統計モデルになる.

3　非線形のモデルに拡張しよう

ところで線形的な関係ではなく, 非線形的だったらどうやって判断すればよいか？

3.1　線形モデルの定義

- Y_i $(1 \le i \le n)$: 反応変数を表す確率変数.
- y_i $(1 \le i \le n)$: Y_i の観測値.
- $(\boldsymbol{x}_i, y_i)$ $(1 \le i \le n)$: データセット. $\boldsymbol{x}_i$ は y_i に関連づいた観測値.

線形モデル

以下の統計モデルを線形モデルという.

$$Y_i = \mu_i + \varepsilon_i, \quad \mu_i = E(Y_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

通常 ε_i は正規分布に従うとする .

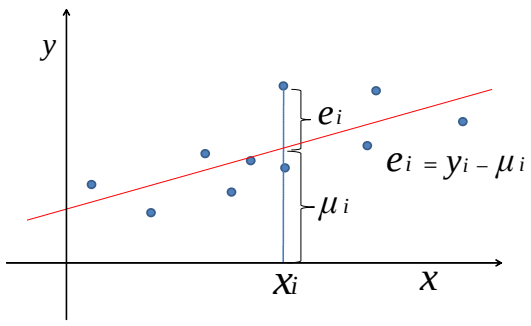


図 1　$\boldsymbol{x}$ が 1 次元の線形モデルのイメージ図

3.2　一般化線形モデルの定義

- Y_i: 反応変数を表す確率変数.
- μ_i: $\mu_i = E(Y_i)$ を満たす予測子.
- $\boldsymbol{x}_i$: Y_i の観測値 y_i に関連づいた観測値.
- g: 予測子の非線形変換 (リンク関数という).

一般化線形モデル

以下の統計モデルを一般化線形モデルという.

$$Y_i = \mu_i + \varepsilon_i, \quad g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

ε_i は指数型分布族と呼ばれる分布に従うとする .

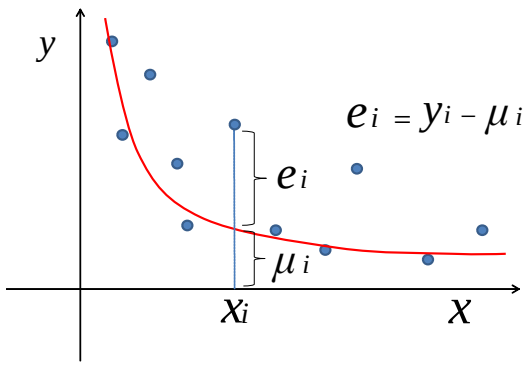


図 2　$\boldsymbol{x}$ が 1 次元の一般化線形モデルのイメージ図

線形モデルは一般化線形モデルの一種である. 他にもロジスティック回帰モデルやポアソン回帰モデル等が存在する.

例：ポアソン回帰モデル

$\boldsymbol{x}$ に対する確率変数を $\boldsymbol{X}$ としたとき,

$$\log(\lambda) = \log(E(Y|\boldsymbol{X})) = \alpha + \boldsymbol{\beta}^\top \boldsymbol{X},$$

$$P(Y = k|\boldsymbol{X}) = \frac{e^{-\lambda}\lambda^k}{k!} \quad (k = 0, 1, 2, \ldots).$$

4　取り組んでいるテーマ

一般化線形モデルに対するモデル評価として Zheng and Agresti [3] による RCC と呼ばれる評価尺度や, エントロピーと呼ばれる尺度を用いた ECD [1] の研究に取り組んでおります → [2].

参考文献

[1] N. Eshima and M. Tabata, *Entropy coefficient of determination for generalized linear models*, Computational Statistics and Data Analysis **54** (2010), 1381–1389.
[2] T. Kurosawa, F.K.C. Hui, A. H. Welsh, K. Shinmura, and N. Eshima, *On goodness-of-fit measures for Poisson regression models*, Australian & New Zealand Journal of Statistics **62** (2020), no. 3, 340–366.
[3] B. Zheng and A. Agresti, *Summarizing the predictive power of a generalized linear model*, Statistics in Medicine **19** (2000), 1771–1781.