医療職のための統計シリーズ

医療職のための学び直し―研究デザインから論文報告までの生物統計学の道標―
第12回　イベント発症リスクに対する回帰モデル

篠崎（シノザキ）　智大（トモヒロ）*

Ⅰ　Look-back：回帰と回帰モデル

回帰（regression）とは条件付き期待値（サブグループ平均値）

$E[\text{結果変数} \mid \text{説明変数} = x]$

のことで，「説明変数＝x」であるようなサブグループの関数$f(x)$として見ることができた。回帰モデル（regression models）は，この$f(x)$を簡略化した数式でまとめて表した近似表現である。例えば直線式

$$f(x) = a + bx$$

を仮定するとき，それは$f(x)$がサブグループxに対して何でも好きな値をとるわけではなく，xに対して直線的にしか変化し得ないという制約を与えている。この制約下で，サブグループxごとの結果変数の平均値（回帰）を，データの平均値になるべく近くなるように求める手法が回帰分析（regression analysis）であった。

上の説明[1)-3)]を手に取るように理解いただけた読者は，前回（**連載第11回**）の内容をしっかり押さえられている。そうでない方は，いま一度前回記事を読み返された上で今回の内容に進んでいただきたい。

Ⅱ　2値結果変数の回帰モデル

（1）　2値結果変数の回帰

過去の連載によると，平均値が適切な要約となる変数は「左右対称に分布する連続変数」であった（**連載第5回**）。ということは，回帰においてサブグループごとの平均値を考えられる「結果変数」は，血圧値などの連続変数である必要があるのだろうか。実はそうではない。医学研究では，例えば「10年以内の動脈硬化症の発症の有無」などの2値変数を数えることが多い。「血圧の平均値」などと異なり「発症有無の平均値」という表現は一見ナンセンスに思えるが「発症あり」＝1「なし」を0という**指示変数**あるいは**インディケータ**で表し直すと（**表1**），この平均値は

$$\frac{1 \times \left(\begin{matrix}\text{発症あり}\\ \text{人数}\end{matrix}\right) + 0 \times \left(\begin{matrix}\text{発症なし}\\ \text{人数}\end{matrix}\right)}{\text{合計人数}} = \frac{\begin{matrix}\text{発症あり}\\ \text{人数}\end{matrix}}{\text{合計人数}}$$

となり，発症割合（incidence proportion）あるいは疾患発症リスク（incidence risk）に一致する（**連載第5回**）。つまり，結果変数が2値の場合にも「回帰」は定義を変えることなく，「条件付き期待値」として「サブグループごとの割合」を表すのに用いることができる：

$$E\left(\begin{matrix}\text{動脈硬化症の}\\ \text{発症インディケータ}\end{matrix} \;\middle|\; \text{年齢} = x \text{歳}\right)$$

$$= P\left(\begin{matrix}\text{動脈硬化症の}\\ \text{発症あり}\end{matrix} \;\middle|\; \text{年齢} = x \text{歳}\right)$$

ここで「P」は括弧内のイベントを生じる確率（probability）すなわち割合・リスクを，縦棒（｜）はやはり条件付けを表す。つまり右辺は「年齢＝x歳」のサブグループでの動脈硬化症の発症リスクを表す**条件付き確率**となる。

回帰の考え方は前回の連続変数の場合と全く

表1　年齢と動脈硬化症発症データの例

患者（i）	年齢（x_i）	動脈硬化症発症	発症インディケータ（y_i）
1	37	なし	0
2	45	あり	1
3	56	なし	0
4	51	なし	0
5	28	なし	0
6	39	なし	0
7	64	なし	0
8	51	なし	0
9	67	あり	1
10	89	なし	0

*東京理科大学工学部情報工学科講師