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11回　回帰モデル

篠崎（シノザキ）　智大（トモヒロ）*

Ⅰ　はじめに

血圧が年齢とともに上昇することはよく知られている（図１）。このような傾向を統計学では

収縮期血圧＝a＋b年齢

という等式で表すことが多い。これは中学校で習う一次方程式なので，統計学の知識がなくても「血圧と年齢が直線的に変化する」こと，「年齢１歳あたり収縮期血圧が$b$mmHg上昇する」ことを読み取るのは難しくないだろう。

しかし，上の等式のように血圧値が年齢で正確に表せるわけではない（同じ年齢でも血圧の値は人によって違うし，同じ人でも１年の血圧上昇幅は異なる。そもそも血圧の値は常に変動する）。そこで，少しお堅い教科書では血圧と年齢の個々の「データ」を直接数式で関係づけるということをする。例えば，患者iの収縮期血圧値をy_i，同じ患者iの年齢をx_iとして測定したとする（表１）。x_iとy_iの散布図（**連載第６回**）を描いてみれば明らかだが，表１のデータは一直線上には乗らないので，ぴったり「収縮期血圧（y_i）＝a＋b年齢（x_i）」となるaとbを選ぶことはできない。そこで患者ごとに血圧の「誤差」と呼ばれるe_iを使って直線からのずれを表すことにして

$$y_i = a + bx_i + e_i$$

という等式を用いる。この数式の適当なaとbを決めるために，x_iを固定して，e_iは互いに独立で平均０かつ分散一定の確率分布に従うと仮定して…教科書ではこのように代数や確率統計を駆使して**回帰分析**（regression analysis）の説明が展開される。

こうした教科書的な説明は実に数学的で，回帰分析の基礎から発展まで厳密な説明ができる。しかし筆者の限られた経験からは，このような枠組みで回帰分析を正しく理解しようとするのは，統計家が自分で思っているより混乱を生じやすいようである。例えば，現実のデータ解析で「ある変数を回帰分析のソフトに入れることで考慮した」というお座なりの説明に対して，何がどう「考慮」されているのかの納得のいく説明を上の枠組みでできる人は多くない。そこで本稿では，以下に述べる「目的に応じた回帰分析の使い方」や，次回扱う「結果変数の型に応じた回帰モデル」を整理して理解する上で，多くの教科書では採用されていない見方から回帰分析の説明[1)2)]を試みたい。

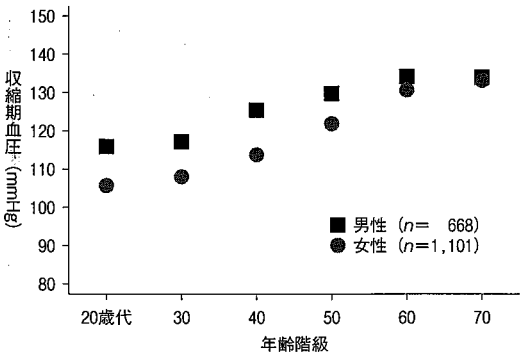

図１　令和元年国民健康・栄養調査による性・年代別収縮期血圧の男女別平均値

表１　収縮期血圧（y_i）と年齢（x_i）のデータ例

患者（i）	年齢（x_i）	収縮期血圧（y_i）
1	37	77.8
2	45	139.9
3	56	114.1
4	51	83.2
5	28	106.7
6	39	69.2
7	64	118.7
8	51	76.7
9	67	146.8
10	89	223.6

＊東京理科大学工学部情報工学科講師