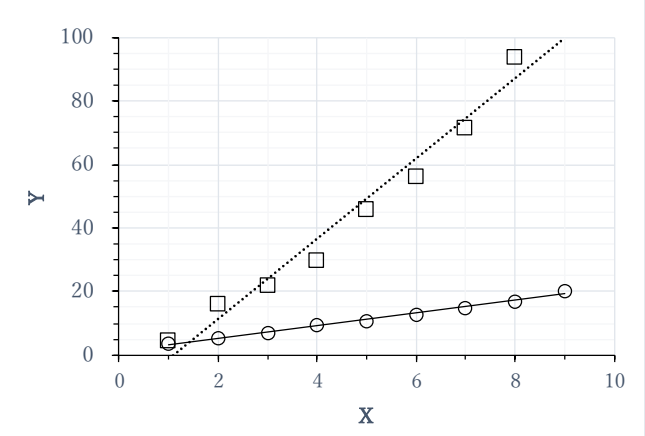
#001★☆☆☆ 2020.3.3 / 2023.7.21

回帰分析では必ずしも相関は直線関係とは限らない。データをプロットして眺めるべき。

年齢と腎機能検査値との関係など、2 つの変数 X と Y との関係性を見たい場合がある。その際、何らかの関係性を得るためにグラフ化して回帰分析を試みるが、用いる関係式は必ずしも直線とは限らない。EXCEL 等では直線を用いた回帰分析が容易に実行できるため、どのようなデータでも直線で関連付けてしまうことがあるが、データを整理したら必ず生データをプロットして眺める習慣をつけるべきである。

下表のデータを EXCEL で入力し、右図のプロットを作成してみる。図にはそれぞれのデータについて直線回帰した結果も描いている。

X	Y1	Y2
1	3.5	4.6
2	5.1	16.0
3	6.9	22.0
4	9.4	29.5
5	10.8	45.6
6	12.9	56.3
7	14.6	71.3
8	17.0	93.7
9	20.0	104.2



図をみると、なんとなくうまく回帰直線が得られたように思われるが、□の図はすこし中央がへこんでいるようにも思われる。こういったことを実際に個々のデータをプロットしてみて人間の感覚で感じ取ることが大切である。□データについて、もう少し詳しく EXCEL の「近似曲線の追加」を用いて解析してみると、下左図のように二次関数のほうがより適切であるような結果が得られ、ここではその違いは小さいが、データをよく見て回帰に用いる式を選ぶことが大切である。

下右図は、たくさんのデータをプロットして「とりあえず」直線回帰で解析した例である。プロットをよくみるとXが大きい場合にYのばらつきも大きく、また指数関数的な増加をしているようにも見える。ではどうすればよいか、「個々のデータをよく見て、数式を正しく使って、自然現象をきちんと理解すること」がデータサイエンスの基本である。