#004★★☆☆ 2020.3.18 / 2023.7.21

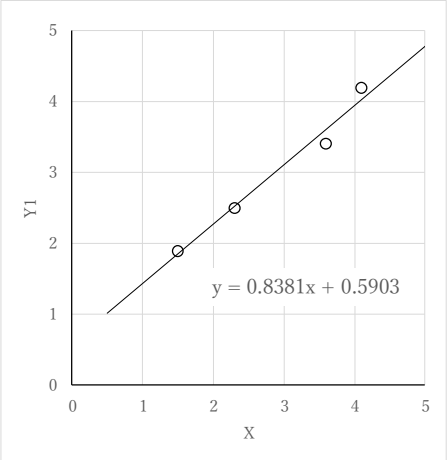
モデル解析の基本～線形回帰分析と最小二乗法の基本原理、相関係数。

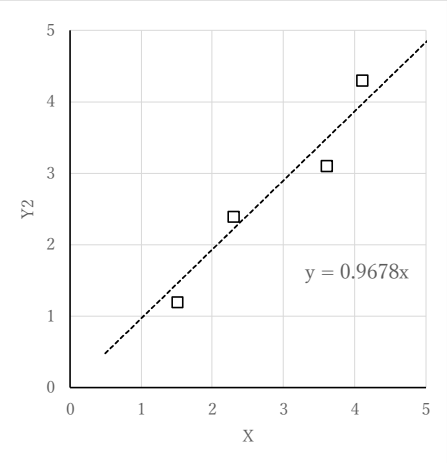
データの傾向や相関性を知るためには「データ全体の真ん中を通る線」を考えると便利な場合が多い。この「線」は必ずしも直線とは限らないが直線だと直感的に理解しやすいため、その数学的手法の「最小二乗法」という方法の説明には直線を用いた回帰分析、すなわち線形回帰分析の例が多く用いられる。

下表のデータを EXCEL で入力し、図のような個々のデータのプロットを作成し、また EXCEL のグラフツールで「近似曲線の追加（線形近似）」の結果を図示してみた。回帰分析を行った結果として図中に数式が得られており、これらの式は最小二乗法という規則に基づいて計算されている。なお、右図は原点を通る直線（切片が 0 の直線）を用いて解析しており、EXCEL であらかじめ切片が 0 となるように設定が必要であることに注意する。

	Original Data	
	X	Y1
1	1.5	1.9
2	2.3	2.5
3	3.6	3.4
4	4.1	4.2

	Original Data	
	X	Y2
1	1.5	1.2
2	2.3	2.4
3	3.6	3.1
4	4.1	4.3





最小二乗法について最も簡単な式として原点を通る（切片が 0 の）直線を考える（右図)。このとき式は $y = \beta$（傾き）$\times\ x$ となり回帰分析では β の値を具体的に求めることになる。右図にある 4 点の座標（x_i, y_i）は表のとおりで（データを区別するために添字をつけた)、傾きが β のときそれぞれの点の「理論的な直線上の」y の値は $\beta \times x$ となる。このときに実測値と理論値との差は（$y - \beta \times x$）となりこれを「残差」と呼び、「真ん中を通る」ということはこの残差がすべての点での和として「最小になる」ことが望ましいことになる。しかしながらこの残差は、実測値と理論値との大小関係によって正の数であったり負の数であったりし、みかけ上全部を足すと「正負の数が打ち消しあってしまう」恐れがある。そこで、すべての残差を二乗して、さらにすべての点について足し合わせた値「残差平方和」が最小となるようにする。これが「最小二乗法」という言葉の由来であり、最小二乗法の理論の背景にはデータが正規分布していると仮定できること、があるが今回は詳しくは触れない。

数式を展開してみる。原点を通る直線について残差平方和は次のようになる。ここで n はデータ数、i はデータポイントそれぞれを指す。

$$SS = \sum_{i=1}^{n} (y_i - \beta \cdot x_i)^2$$

この式について SS が最小となる β を得るためには SS を β で（偏）微分した次の式が 0 となることが必

要条件となる。

$$\frac{\partial SS}{\partial \beta} = 2 \times (-x_i) \sum_{i=1}^{n} (y_i - \beta \cdot x_i) = 2 \sum_{i=1}^{n} (\beta \cdot x_i \cdot x_i - x_i \cdot y_i) = 0$$

これを解くと次のように β の値が実測値から計算できる（十分条件の検証については省略)。

$$\beta = \sum_{i=1}^{n} \frac{x_i \cdot y_i}{x_i \cdot x_i}$$

切片のある直線を用いるような場合も線形回帰分析のひとつであるが、切片や傾きを求めるための数式はかなり複雑になり手計算で求めずに上記 EXCEL のようにコンピュータを使った計算で求める。以下に EXCEL の「分析ツール」で切片がある場合の回帰分析（グラフ左側）を行った結果を示す。具体的な計算方法は EXCEL の操作ガイドに従ってほしい。

概要								
回帰統計								
重相関 R	0.987467							
重決定 R2	0.975091076							
補正 R2	0.962636614							
標準誤差	0.1952195							
観測数	4							
分散分析表								
	自由度	変動	分散	測された分散	有意 F			
回帰	1	2.983778693	2.983778693	78.29250965	0.012533			
残差	2	0.076221307	0.038110653					
合計	3	3.06						
	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	0.590347263	0.289293804	2.040649524	0.178082394	-0.65438351	1.835078038	-0.65438351	1.835078038
X 値 1	0.838140082	0.094723231	8.848305468	0.012533	0.430578914	1.245701251	0.430578914	1.245701251

「重相関 R」は相関係数を指し -1 から+1 の値をとり、負数の場合には負の相関、すなわち X が増えると Y が減る、正数の場合には正の相関となる。相関係数の絶対値が 1 に近いほど「強い相関」となる。表の下方に「係数」とあってその値が推定された切片および傾き（X 値 1）で上図中の値と同じであることが確認できる。これらの推定値の統計的有意性を「P-値」で確認でき、この例では切片の p 値は一般的に有意水準とされる 0.05 よりも大きいので「切片は有意な値（0 と有意に異なる値）ではなく、0 とみなすことが統計的に適切」と判断できる。一方傾きについて p 値は 0.05 よりも小さく「傾きは統計的に有意である（傾きは 0 とは有意に異なる値である)」と判断できる。

このように回帰分析の有意性については、切片や傾きなどの推定するパラメータの統計的有意性をもとに判断するべきであり、相関係数の値だけでは「強い相関」「弱い相関」といった相対的な評価はできるが統計的観点からは十分ではない。回帰分析を行った場合には、回帰直線と実測値のプロットで視覚的に妥当性を確認した上で、それぞれの係数の統計的有意性をみて議論を行うことが大切である。