回帰分析

小テストはゴムの自然長を変え、1N の力で引っ張った時の長さのデータを解析した。その際、自然長と伸びた後の長さの散布図を書くと明らかに線形の関係が認められ散布図に近似曲線として 1 次関数を追加すると非常に正確にデータを再現することが確認できた。

このような近似関数を求めることにより、自然長を測定していないゴムでも伸びた後の長さを予想できたりするようになる。

科学は、このように実験を行ない、その事象の原因と結果を結びつけることを基本とする。言い替えると、事象 A と事象 B の間の因果関係を明らかにすることである。

ところが、因果関係を明らかにすることは容易ではない。それどころか、ある結果をもたらす要因が明らかではないことも多い。

相関関係と因果関係

科学では因果関係を明らかにすることを目的とするが、因果関係を調べるためには大きく分けて 2 つの方法がある。数学の命題の証明を思い出してみると良いかも知れない...

1. 因果関係を否定するためには、反証を示す
2. 因果関係を肯定するには、あらゆる条件で、その原因事象から結果の事象に繋がる規則性などの関係が成立することを検証する

数学の命題とは異なり、肯定するためには、2 つ (以上の) 事象の関連を調べる必要があり、統計処理を用いた相関関係の 検証 が必要となる。

相関関係は因果関係と異なり 2 つの事象の間における、原因と結果のつながりの有無については言及しない。したがって、相関関係が認められた事象 AB は次のように分類される。

1. A → B の因果関係が認められる場合
2. B → A の因果関係が認められる場合
3. C → A, C → B という共通原因がある場合
4. 偶然。(じつは A と B にはなんの関係もない)

このように因果関係と相関関係は異なるものであるということを認識した上で相関関係について学習をする

相関関係を確かめるために

相関関係を確かめるためには、結果の事象を表す変数 (従属変数という) と原因を表す変数 (独立変数という) の間に一定の規則があることを仮定し、その仮定が正しいか否かの判断を行なう必要がある。

その判断を行なう方法の一つに回帰分析がある。回帰分析とは、従属変数が独立変数からなる計算式(多くの場合は線形) で表されると仮定し、その計算式に現れるパラメタを決定する分析方法である。

従属変数は常に一つだが、独立変数は単数でも複数でも良く、一つの場合は特に「単回帰」、複数の場合「重回帰」と呼ぶ事もある。また求められた計算式を回帰式と呼ぶ。

一般に、回帰式は線形であるが、パラメタ変換 (対数をとる、指数をとる、逆数をとるなど) を行なって線形に帰着できる、指数関数や対数関数への拡張は非常に単純である。

最小二乗法による単回帰分析の例

小テストと同様な実験を再度行なったこととして、新たなデータを分析してみよう。

測定されたデータの組を (x_i, y_i) と表すこととする。

この組をもっともうまく説明する一次関数を決定するということは、求める一次関数を

$$y \quad = \quad ax + b \tag{1}$$

とおいた上で「もっともうまく説明する」 a と b を決定するということである。一般的な回帰分析では最小二乗法という、目的直線とデータポイントのずれの二乗を最小にするような方法が使われる。

つまり、x_i が与えられた際の予想される値 $(ax_i + b)$ と y_i の差の二乗の総和を計算し、それを最小とする (a, b) を決定するのである。

最小にするべきものを具体的に書き下すと

$$\sum_{i=1}^{n}(y_i - (ax_i + b))^2 \tag{2}$$

となり、この値を最小とする 一次関数の係数 a と切片 b を求めるのである。この式に現れる項のうち、すべての x_i および y_i は既知であるので、この式に現れる未知のものは a, b だけであることに注意する。

したがって、

$$F(a,b) = \sum_{i=1}^{n}(y_i - (ax_i + b))^2 \tag{3}$$

と a, b の関数として定義し、$F(a,b)$ の極小値を探索する事になる。

高校数学で学んだように連続関数の極小値を探索するためには、微分係数が 0 になる点を探せば良いのであった。

したがって、極小点では、以下の 2 式が成立する。

$$\frac{\partial F}{\partial a} = -2\sum_{i=1}^{n}(y_i - (ax_i + b))x_i = 0 \tag{4}$$

$$\frac{\partial F}{\partial b} = -2\sum_{i=1}^{n}(y_i - (ax_i + b)) = 0 \tag{5}$$

式 (5) より

$$\sum_{i=1}^{n} y_i = a\sum_{i=1}^{n} x_i + nb \tag{6}$$

となり、両辺を n で割る事で

$$\frac{1}{n}\sum_{i=1}^{n} y_i = a\frac{1}{n}\sum_{i=1}^{n} x_i + b \tag{7}$$

$$\bar{y} = a\bar{x} + b \tag{8}$$

となるので

$$b = \bar{y} - a\bar{x} \tag{9}$$

を得る。

方程式 (4), (5) に代入する事で

$$\sum_{i=1}^{n}\{(y_i-\bar{y})-a(x_i-\bar{x})\}x_i=0 \tag{10}$$

$$\sum_{i=1}^{n}\{(y_i-\bar{y})-a(x_i-\bar{x})\}=0 \tag{11}$$

が得られ、式 (11) に $\bar{x}$ を掛けたものを (10) から差し引く事で次式を得る。

$$\sum_{i=1}^{n}\{(y_i-\bar{y})(x_i-\bar{x})\}=a\sum_{i=1}^{n}\{(x_i-\bar{x})^2\} \tag{12}$$

したがって、

$$a = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{nS_{xy}}{nS_{xx}}=\frac{S_{xy}}{S_{xx}} \tag{13}$$

$$b = \bar{y}-\frac{S_{xy}}{S_{xx}}\bar{x} \tag{14}$$

と表す事ができる。ただし、

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i \tag{15}$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n}y_i \tag{16}$$

$$S_{xx} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2 \tag{17}$$

$$S_{yy} = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2=\frac{1}{n}\sum_{i=1}^{n}y_i^2-\bar{y}^2 \tag{18}$$

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\bar{x}\bar{y} \tag{19}$$

である。

なお、方程式 (4), (5) を直接解けば

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \tag{20}$$

$$b = \frac{-\sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \tag{21}$$

という形式の解を得る。

相関係数

回帰分析ができたとして、その得られた式がもっともらしいかどうかの評価は別に行なわなければならない。

例えば、次に示すグラフは測定されたデータを元に最小二乗法に基づいて回帰直線を求めたものであるが、明らかにもっともらしさは異なる。

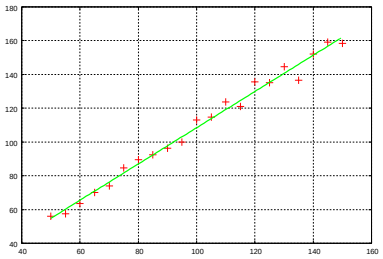

図 1: $R^2 = 0.9885$

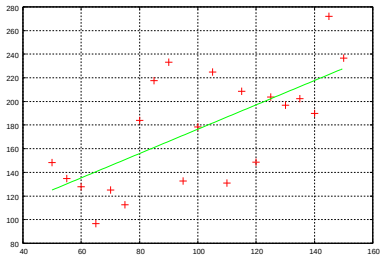

図 2: $R^2 = 0.4439$

これらを評価するための指標として、相関係数を次のような式に基づいて導入する

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{22}$$

これは、n 次元ベクトル $\vec{x} - \bar{\vec{x}}$ と $\vec{y} - \bar{\vec{y}}$ のなす角の余弦 (cos) に等しい。したがって、相関係数は必ず -1 と 1 の間の値をとる。

この相関係数をもとに二つの指標の線形関係の強さの評価を行なう。

相関係数が 0.7 以上あれば、正の相関 (片方の値が上がれば、もう片方の値も一定の割合で上がる関係) があるとみなされ、逆に、-0.7 以下であれば、負の相関 (片方の値が上がれば、もう片方の値が一定の割合で下がる関係) があるとみなされる。

相関係数の絶対値が 1 の場合、二つの変数は線形従属の関係であると結論づけられることになる。

最小二乗法による重回帰分析

目的変数を説明すると考えられる独立変数が複数となった場合でも、考える方法は同様である。二つの独立変数 (x, y) から目的変数 (z) を説明する場合は式 (1) の変わりに

$$z = ax + by + c \tag{23}$$

を仮定し、おなじ事を行なう。変数が一つ増えるが、偏微分 $= 0$ の式も一つ増えるため、同じ条件で a, b, c を決定する事が可能である。

独立変数が 3 以上となっても手順は同様である。

単回帰と大きく異なる点は、独立変数として取り扱う変数が「実際に独立」であるかどうかの評価が必要となる点である。したがって、独立変数間での相関関係を求め、相関が高くない ように変数を選択しなければならない。