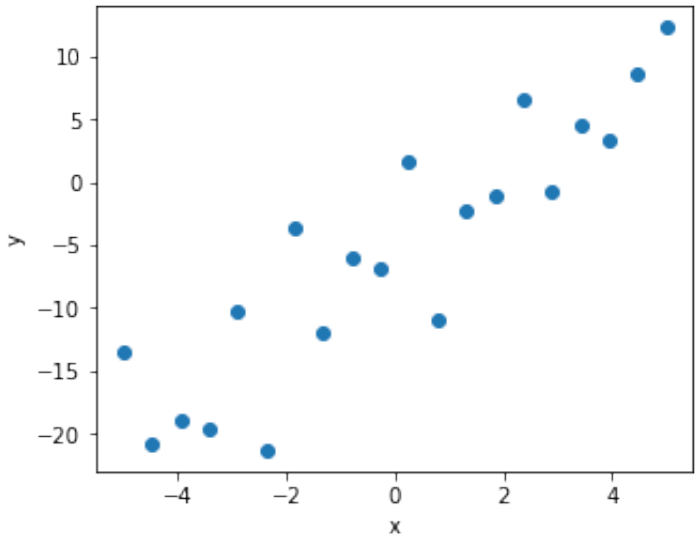
線形相関

条件をいろいろ変えながら、実験を行い、測定すると、条件にあわせて測定値が変動します。このような、条件(=入力)と測定結果(=出力)の間にもし直線的な相関が見付けられれば、すべての条件で実験を行わなくても、結果が予想できるようになります。グラフ用紙に、実験条件を横軸に、実験結果を縦軸にとって、入力と出力の対応をプロットすると、散布図が得られますが、散布図が直線的に広がっているなら、直線で近似するのはまあ妥当と考えられるでしょう。



この時、直線はどのように引けば一番もっともらしいと言えるでしょうか。

ここでも、線の引き方にはいろいろ選択肢がありますが、ここでは、一番計算が簡単になるようなケースだけを説明します。まず、横軸にとる実験条件の数値x_iの測定精度は、縦軸の実験値y_iの測定精度に比べて十分に高いこと、そして、実験条件x_iの値によらず、実験値y_iは同じ程度にばらついていることを仮定します。そして、仮に直線$y = ax + b$を引いた時に、x_iでの直線上の点$y'_i = ax_i + b$と、実際の実験値y_iの違いが一番小さくなるように、直線の傾きaと切片bを選ぶのです。

実験データの点の分布と、仮に引いた直線とのずれの大きさを次の量Eではかることにします。

$$E = \sum_{i=1}^{N} (y'_i - y_i)^2$$

二乗をつけてあるのは、直線に対して実験値が上にある場合も下にある場合もあり、ずれが大きいほど、より大きな値にあるようにしたいからです。(もちろん、ほかの関数を使うこともできますが、あとの計算が少し面倒になります。)

E が極小値をとるような a と b は、Eを a と b で偏微分することで求められます。

$$\frac{\partial E}{\partial a} = 0$$

$$\frac{\partial E}{\partial b} = 0$$

偏微分を実行する前に、Eを展開しておきます。

$$E = \sum_{i=1}^{N} (y'_i - y_i)^2$$

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

$$= \sum_{i=1}^{N} (a^2 x_i^2 + b^2 + y_i^2 + 2abx_i - 2by_i - 2ax_i y_i)$$

$$= a^2 \sum_{i=1}^{N} x_i^2 + Nb^2 + \sum_{i=1}^{N} y_i^2 + 2ab \sum_{i=1}^{N} x_i - 2b \sum_{i=1}^{N} y_i - 2a \sum_{i=1}^{N} x_i y_i$$

面倒なので、x_i の二乗和$\sum_{i=1}^{N} x_i^2$を、$[x^2]$などと書くことにすると、

$$E = a^2 [x^2] + Nb^2 + [y^2] + 2ab[x] - 2b[y] - 2a[xy]$$

と書けます。ここまで展開しておけば、aとbで偏微分するのは簡単です。

$$\frac{\partial E}{\partial a} = 2a[x^2] + 2b[x] - 2[xy] = 0$$

$$\frac{\partial E}{\partial b} = 2Nb + 2a[x] - 2[y] = 0$$

つまり、直線の切片bと傾きaは、次の5つの量を計算すれば、簡単に決まります。

1. 実験条件 x_i の総和 $[x]$
2. 実験条件 x_i の二乗の総和 $[x^2]$
3. 実験値 y_i の総和 $[y]$
4. 実験条件 x_i と実験値 y_i の積の総和 $[xy]$
5. 実験値の個数 N

このように、ちらばったデータを、直線で近似することを線形回帰(linear regression)と呼びます。直線ではなく二次曲線や、その他一般的な曲線で回帰する場合も、手続きは全く同じです。ただ、回帰する関数の変数が多くなると、未知パラメータが増え、偏微分の式の項数が増えるので、連立方程式を解くのが多少難しくなります。

直線で回帰する場合でも、その直線の傾きがあらかじめわかっている場合には、a は定数になりますから、解くべき微分方程式は1つだけになります。切片があらかじめわかっている場合にも同様です。例えば、一定量の気体の、いくつかの温度での、圧力の実測値のデータを得て、それをもとに気体の分子数を逆算する問題を考えます。シャルルの法則によれば、圧力は絶対温度に比例するはずですから、横軸に絶対温度、縦軸に圧力をとれば、回帰直線は原点(切片0)を通らなければなりません。$b = 0$を代入した上で、Eを直線の傾き a で偏微分すると、

$$\frac{\partial E}{\partial a} = 2a[x^2] - 2[xy] = 0$$

という簡単な式が1つだけ得られます。つまり、直線の傾きは $a = [xy]/[x^2]$ で求められます。そして、傾きから、アボガドロの法則により分子数が求まります。

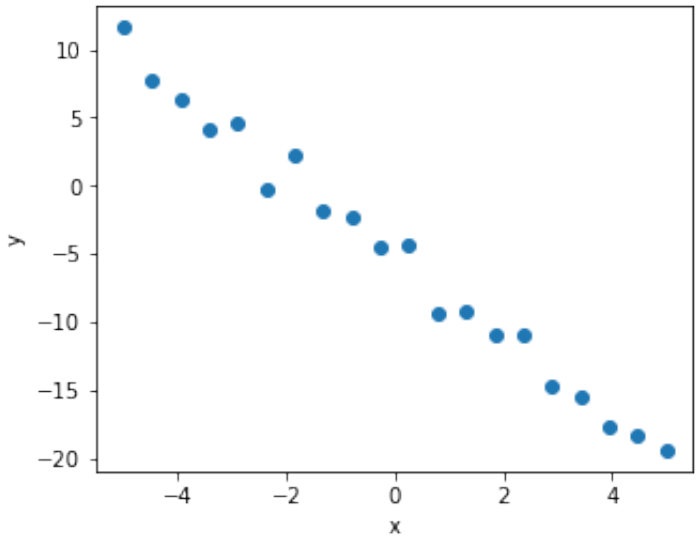
このように、どんな関数で回帰すべきなのか(一般直線でいいのか、原点を通る直線なのか、あるいは二次曲線や指数関数のほうがいいのか)、その場合に、a や b といったパラメータが、回帰で推定しなければいけない量なのか、あらかじめ固定されている量なのかをよく考えてから計算をするようにしましょう。そうでないと、意味不明の統計量が得られることになります。

Pearson相関係数

上の例では、十分信頼できる入力値を横軸に、計測値(出力)を縦軸にとりました。横軸は実験者が意図を持って決めている値なのでコントロールとも呼ばれます。

もうすこし一般的な場合として、データ対の集合が与えられた場合を考えます。この場合も、片方をx軸に、もう片方をy軸にとって、散布図を描くことはできますが、どちらかがもう一方をコントロールしているかもしれないし、全く無関係かもしれません。

それでも、散布図を見れば、データの特徴がつかめます。次の図のように、点のひろがりが小さい場合には、横軸の値を見れば、縦軸の値がおおよそ予測できるでしょうし、逆の予測もそれほど的はずれにはならないでしょう。



この、点のひろがりの大きさを表す尺度がPearson相関係数です。

2つのデータ列A,B(例えば時系列データ)の各要素をa_i,b_iと表すとき，これらの間の共分散(covariance)を次のように定義する。

$$f(A,B) = \frac{1}{N}\sum_{i=1}^{N}(a_i - \mu_A)(b_i - \mu_B)$$

μ_AはAの平均値。この式では、fは2つの座標の関数であり、相関関数と呼ばれることもある。 共分散は、a_iとb_iが同じように増減すると、絶対値が大きくなる。取り扱いやすいように、共分散を，Aの標準偏差

$$\sigma_A = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(a_i - \mu_A)^2}$$

と σ_B で割って規格化したものをPearson相関係数と呼ぶ。

$$\rho(A,B) = \frac{f(A,B)}{\sigma_A \sigma_B}$$

Pearson相関係数は$-1 \sim 1$の値をとり、点のひろがりが小さいほど1(正の相関)あるいは-1(負の相関)に近付く。データが完全に直線上にのる場合には+1または-1となる。

Pearson相関係数は、AとBの間に、比例関係が成りたつ(線形相関)ことを仮定している。AとBが関連して変化していたとしても、線形相関でなければ、捉えることができないことに注意。

In [ ]: