第4回　線形回帰と重回帰分析

エクセルが見える位置に着席してください

CES: https://www.ces-alpha.org/hp/DSB2022/top/?timestamp=485211

2022/6/23 (木) DSB

9章　相関と線形回帰

偏差 (1章)：

変数 $x = x_1, x_2, x_3, \ldots, x_n$

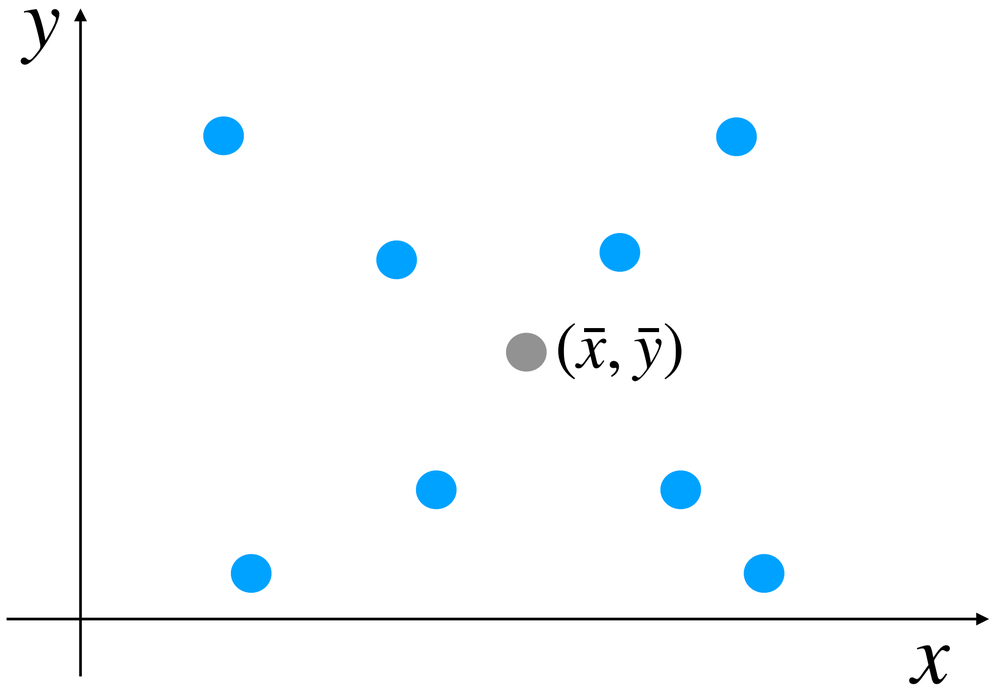
変数 $y = y_1, y_2, y_3, \ldots, y_n$

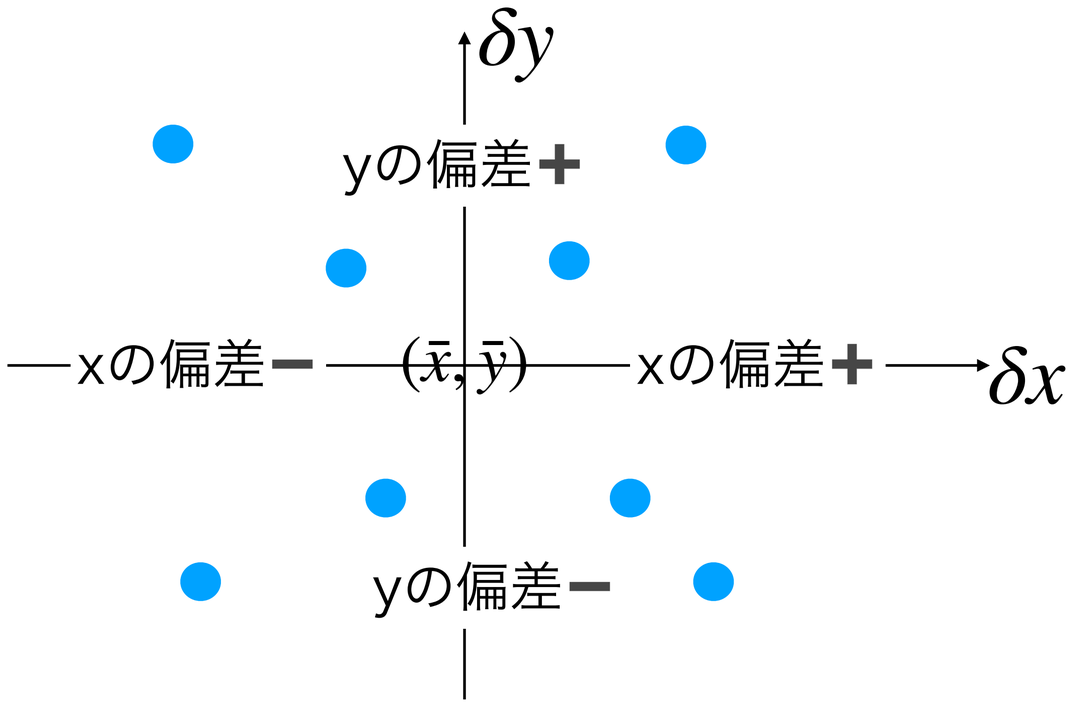
2変数間の関係を見つけて、任意のxに対するyの値を予測したい

偏差:deviation

$$\delta x_i = x_i - \frac{1}{n}\sum_i^n x_i = x_i - \bar{x}$$

$$\delta y_i = y_i - \frac{1}{n}\sum_i^n y_i = y_i - \bar{y}$$





9章まとめ1：共分散

変数 $x = x_1, x_2, x_3, \ldots, x_n$

変数 $y = y_1, y_2, y_3, \ldots, y_n$

偏差:deviation

$$\delta x_i = x_i - \frac{1}{n}\sum_i^n x_i = x_i - \bar{x}$$

$$\delta y_i = y_i - \frac{1}{n}\sum_i^n y_i = y_i - \bar{y}$$

共分散:covariance

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n} \delta x_i \delta y_i$$

(x_iの偏差 x y_iの偏差)の平均

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} = \overline{xy} - \bar{x}\bar{y}$$

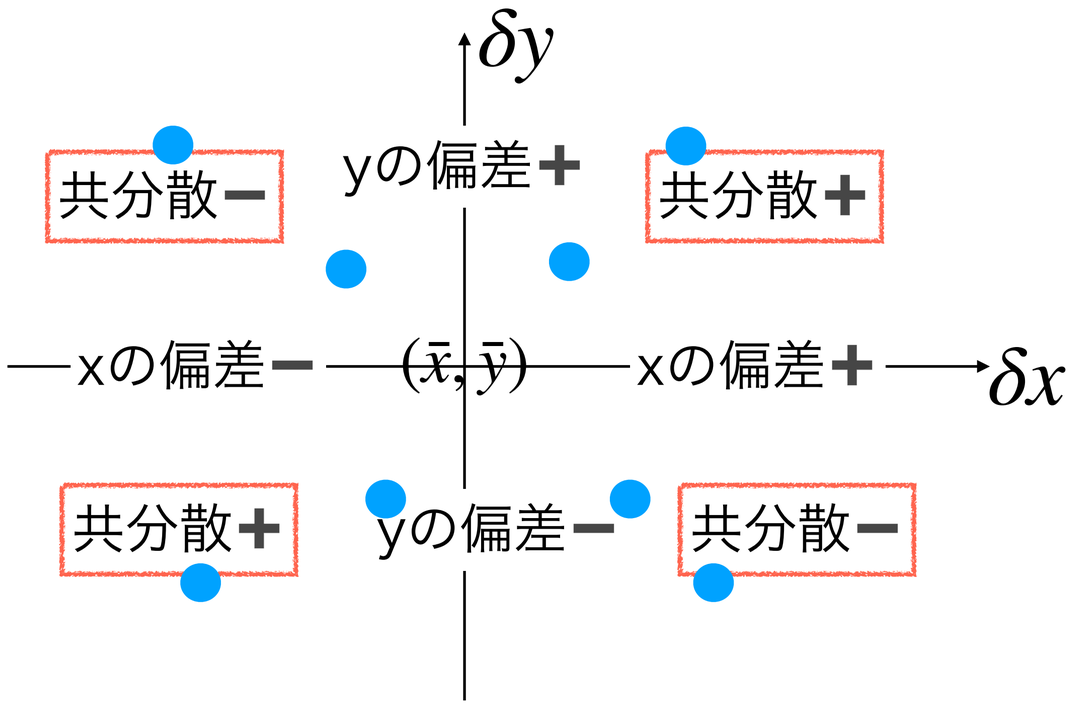
積$x_i y_i$の平均 - 平均の積

$y_i = x_i$のとき

$$\sigma_{xx} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \overline{x^2} - \bar{x}^2 = \sigma_x^2$$

xの２乗の平均 - 平均の２乗

(1.6)xの分散:variance



ランダムなデータの場合、
共分散の期待値(平均)はゼロになる

9章まとめ2：相関係数

2変数間の関係している度合い

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

xyの共分散:covariance

x,yの標準偏差の積

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n} \delta x_i \delta y_i$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} \delta x_i^2$$

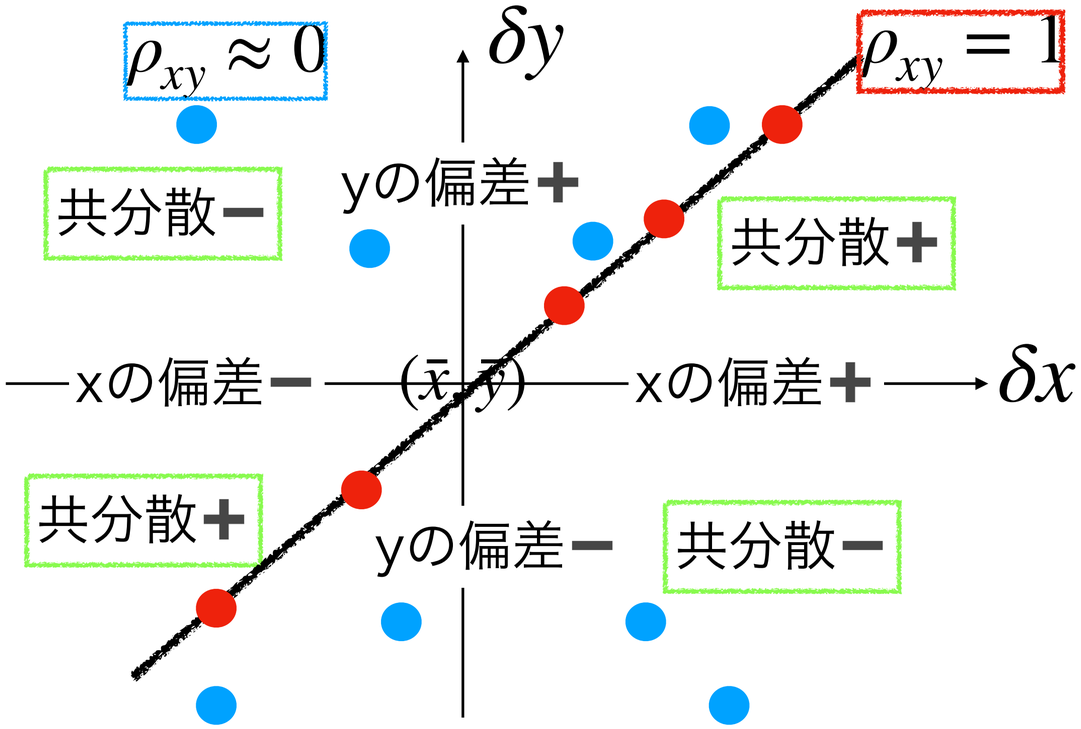
分散のルートが
標準偏差(standard deviation)

$|\rho_{xy}| = 1$：y=ax+bのような完全な相関があるとき

$\rho_{xy} \approx 0$：相関がない

($|\rho_{xy}| \geq 0.9$：強い相関がある)

相関係数の絶対値は0から1の値を取りうる
(右肩下がりの相関がある時、相関係数は負になる)
相関係数にx,yの標準偏差を掛けると共分散になる



9章まとめ3：線形回帰（最小二乗法）

データ点と直線上の値のズレ(残差)の平均

$$L = \frac{1}{n}\sum_{i}^{n}\left[y_i - (ax_i + b)\right]^2$$

が最小になるようなa,bを決めるのが
線形回帰(linear regression)
最小二乗法(least-square method)

$\frac{dL}{da} = 0, \quad \frac{dL}{db} = 0$　を満たすa,bを求めると、

$$a = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\sigma_{xy}}{\sigma_x^2}$$

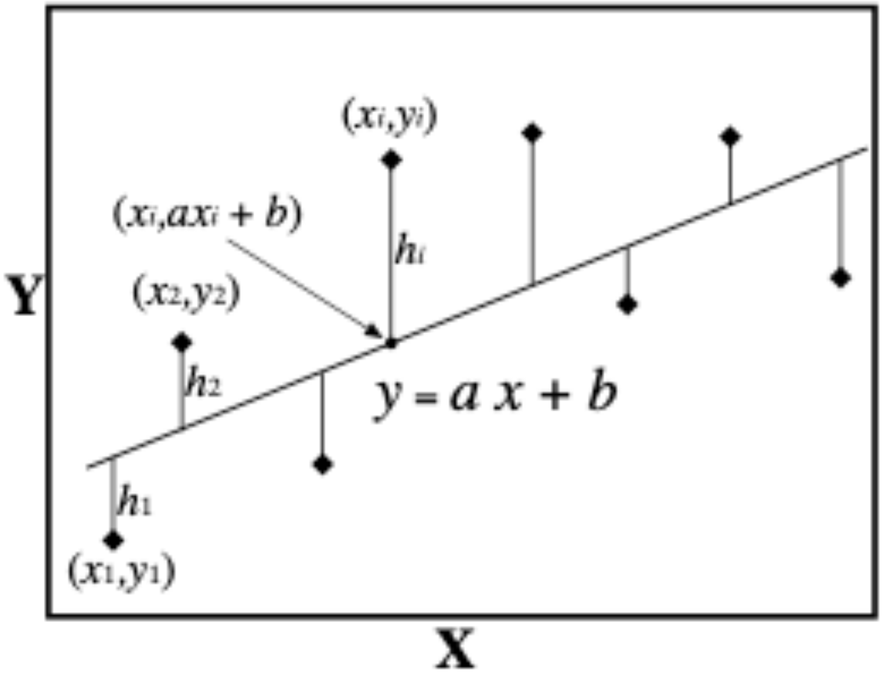
$\frac{\text{xyの共分散}}{\text{xの分散}}$

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

x,yの平均から求められる

導出は先週の概説ノートを参照or付録A.7にも記載されています

データの変化の傾向を表す直線
全てのデータ点を代表する直線　図9.6



線形回帰直線　$y = ax + b$

がわかると、任意のXに対するyを予測できる!

テキスト例題9.2

SUM, AVERAGE
あたりは使ってよい

線形回帰直線 $y = ax + b$

それぞれの項を計算する $\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n} \delta x_i \delta y_i$

成人男性6人の靴のサイズと身長のデータから回帰直線の係数を求めよ.
散布図を作成して直線を描くこと.
エクセル20220623 シート「9.2」参照

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} \delta x_i^2$$

$$a = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\sigma_{xy}}{\sigma_x^2}$$

一気にaを計算するときは、(分子)/(分母)のようにグループとなる計算に括弧をつけましょう。括弧がないと割り算・掛け算が優先されてしまいます。

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

- 平均、分散、標準偏差、共分散、相関係数
- 線形回帰直線の傾きa, 切片b
- 散布図の作成：x, yの配列を選択→挿入→散布図
- 回帰直線の表示：データ点を右クリック→近似曲線の追加→線形近似
  回帰直線を右クリック→近似曲線の書式設定→グラフに数式を表示するに✅

- (参考:エクセルの関数)
  - 分散VAR.P, 標準偏差STDEV.P, 共分散COVARIANCE.P(xの配列,yの配列),
    相関係数 CORREL(xの配列,yの配列)
  - 線形回帰の傾きa: SLOPE(yの配列,xの配列)
    切片b: INTERCEPT(yの配列,xの配列)

補足：エクセルで２乗する時は、例えば、「=B2^2」で２乗になる。それを列全体に適用すれば良い。

演習例題1

SUM, AVERAGE
あたりは使ってよい

線形回帰直線 $y = ap + b$

それぞれの項を計算する $\sigma_{py} = \frac{1}{n}\sum_{i=1}^{n} \delta p_i \delta y_i$

最高気温とお店の客数のデータから回帰直線を求めよ
エクセル20220623 シート「線形回帰」参照
エクセルの関数を使わずに一つずつ確認してみましょう

$$\sigma_p^2 = \frac{1}{n}\sum_{i=1}^{n} \delta p_i^2 \qquad a = \frac{\overline{py} - \bar{p}\bar{y}}{\overline{p^2} - \bar{p}^2} = \frac{\sigma_{py}}{\sigma_p^2}$$

一気にaを計算するときは、(分子)/(分母)のようにグループとなる計算に括弧をつけましょう。括弧がないと割り算・掛け算が優先されてしまいます。

$$b = \bar{y} - a\bar{p}$$

- 平均、分散、標準偏差、共分散、相関係数
- 線形回帰の傾きa, 切片b
- 散布図の作成：p, yの配列を選択→挿入→散布図
- 回帰直線の表示：データ点を右クリック→近似曲線の追加→線形近似
  回帰直線を右クリック→近似曲線の書式設定→グラフに数式を表示するに✅
- (参考:エクセルの関数)
  - 分散VAR.P, 標準偏差STDEV.P, 共分散COVARIANCE.P(pの配列,yの配列),
    相関係数 CORREL(pの配列,yの配列)
  - 線形回帰の傾きa: SLOPE(yの配列,pの配列)
    切片b: INTERCEPT(yの配列,pの配列)

補足：エクセルで２乗する時は、例えば、「=B2^2」で２乗になる。それを列全体に適用すれば良い。

例題2 テキスト章末問題9-1(1-3のみ。4は除く)

8個の卵の全体の質量と黄身の質量のデータから回帰直線を求めよ。

エクセルDSB_20220623 シート「9-1」参照

同じようにエクセルの関数を使わずに平均、分散、標準偏差、a, bの値を求め、回帰直線を求めてみましょう

答えは付録E(p.186)に一応あります

重回帰分析

線形回帰

例：変数p（最高気温）と変数y（客数）

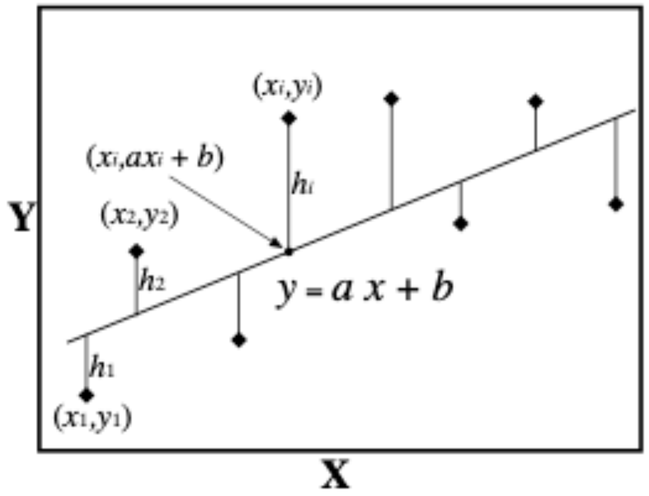
$$p = p_1, p_2, p_3, \ldots, p_n$$

$$y = y_1, y_2, y_3, \ldots, y_n$$

線形回帰直線 $y = ap + b$

a, b:定数 を求める

→任意のpでyを予測



重回帰

プラスもう一つ（以上）の変数

例：変数p（最高気温）と**変数q(最低気温)**と変数y（客数）

yにはpだけでなく
qも影響している？？

$$p = p_1, p_2, p_3, \ldots, p_n$$

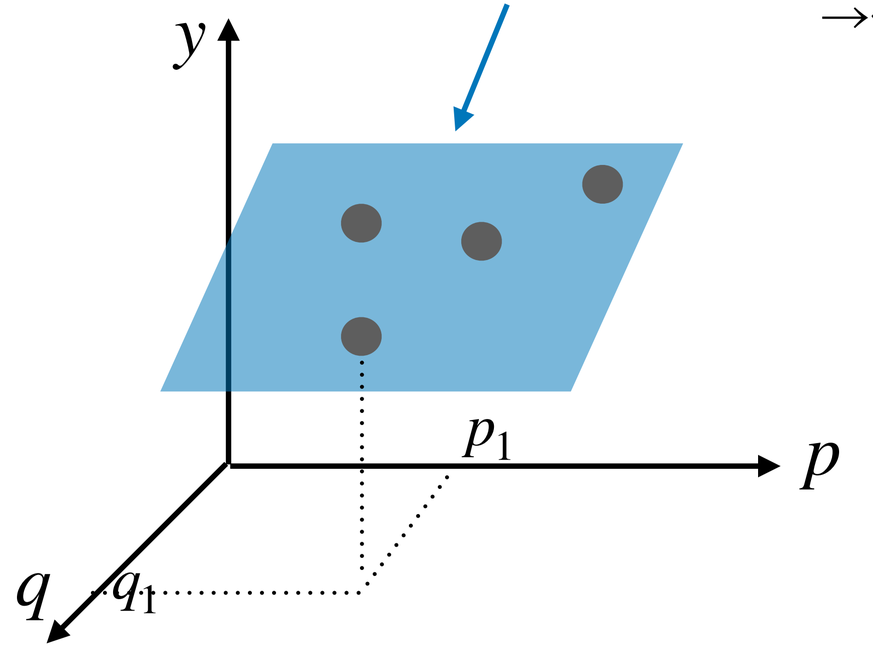
$$q = q_1, q_2, q_3, \ldots, q_n$$

$$y = y_1, y_2, y_3, \ldots, y_n$$

重回帰直線 $y = ep + fq + g$

e, f, g:定数 を求める

→任意のpでyを予測



重回帰

プラスもう一つ（以上）の変数

例：変数p（最高気温）と**変数q(最低気温)**と変数y（客数）

yにはpだけでなくqも影響している？？

$$p = p_1, p_2, p_3, \ldots, p_n$$

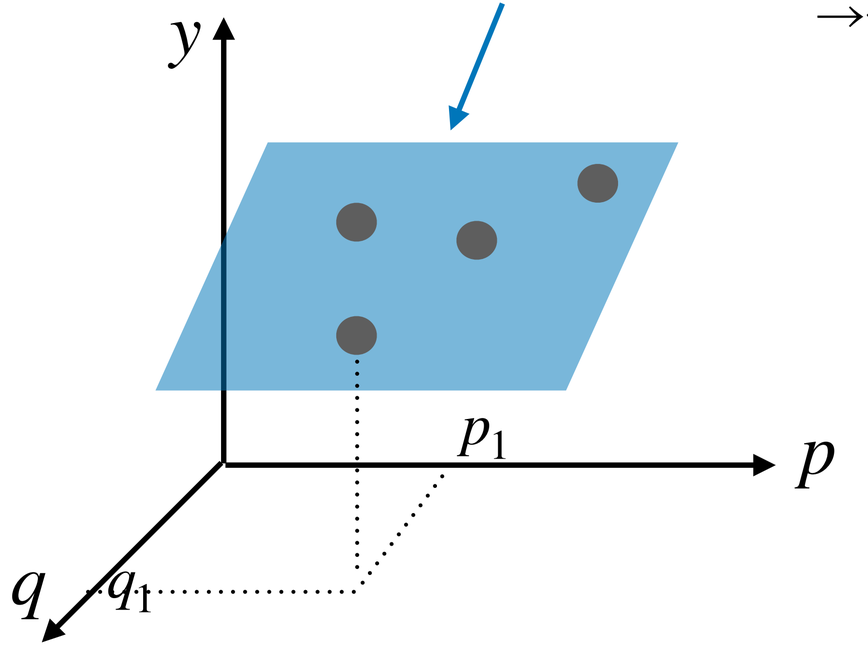
$$q = q_1, q_2, q_3, \ldots, q_n$$

$$y = y_1, y_2, y_3, \ldots, y_n$$

重回帰直線　$y = ep + fq + g$

e, f, g:定数 を求める

→任意のpでyを予測



相関係数

$$\rho_{py} = \frac{\sigma_{py}}{\sigma_p \sigma_y} \quad \rho_{qy} = \frac{\sigma_{qy}}{\sigma_q \sigma_y} \quad \rho_{pq} = \frac{\sigma_{pq}}{\sigma_p \sigma_q}$$

偏相関

pの影響を除いた、qとyの偏相関係数

$$\frac{\rho_{qy} - (\rho_{py} \times \rho_{pq})}{\sqrt{1 - \rho_{py}^2} \times \sqrt{1 - \rho_{pq}^2}}$$

qの影響を除いた、pとyの偏相関係数

$$\frac{\rho_{py} - (\rho_{qy} \times \rho_{pq})}{\sqrt{1 - \rho_{qy}^2} \times \sqrt{1 - \rho_{pq}^2}}$$

重回帰

プラスもう一つ（以上）の変数

例：変数p（最高気温）と**変数q(最低気温)**と変数y（客数）

yにはpだけでなくqも影響している？？

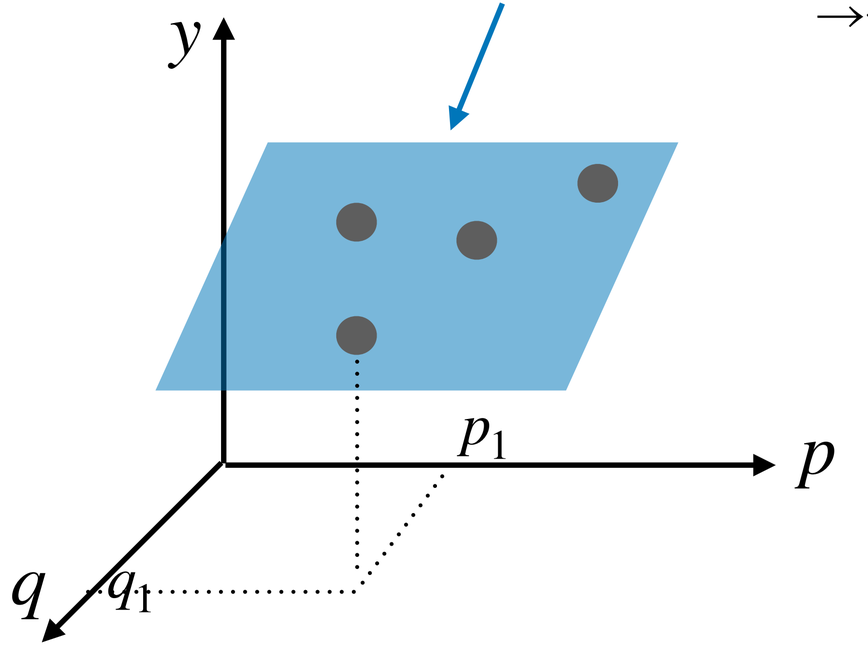
$$p = p_1, p_2, p_3, \ldots, p_n$$

$$q = q_1, q_2, q_3, \ldots, q_n$$

$$y = y_1, y_2, y_3, \ldots, y_n$$

重回帰直線　$y = ep + fq + g$　e, f, g:定数 を求める

→任意のpでyを予測



qの影響を除いた、pとyの偏回帰係数

$$\frac{\rho_{py} - (\rho_{qy} \times \rho_{pq})}{1 - \rho_{pq}^2} \times \frac{\sigma_y}{\sigma_p} = e$$

pの影響を除いた、qとyの偏回帰係数

$$\frac{\rho_{qy} - (\rho_{py} \times \rho_{pq})}{1 - \rho_{pq}^2} \times \frac{\sigma_y}{\sigma_q} = f$$

定数g　$g = \bar{y} - e\bar{p} - f\bar{q}$

エクセル20220623 シート「重回帰」参照

３変数のデータ配列を選択→「データ」タブ

→データ分析→回帰分析→入力x範囲にp~qの配列を入れる

→新しいワークシートに重回帰分析の結果が出る

pythonの場合はscikit-learnで簡単にできるみたいです

重回帰分析

- 客数を最低気温と最高気温の2つの変数から予測する
    - 日照時間、湿度、曜日、、などの変数をさらに足していくと、より正確に客数を予想できそう
- ただ変数を足していけばよいのではなく、どの変数が重要かを、偏相関係数を使ってひとつひとつ調べていくと、一見、目的変数（この場合は客数）との相関は弱くてもより精度の高い予測に役立つ変数を探すことができる
- 例えば、生ビールの売り上げとアイスクリームの売り上げの相関が強いからといって、2つの間に直接の関係がある（アイスクリームが売れたから生ビールが売れた）とは限らない。実際、最高気温と生ビール売り上げ、最高気温とアイスクリーム売り上げ、は真の相関があったとすると、最高気温の影響を除いた双方の売り上げの偏相関係数を求めると、その値は小さくなる。