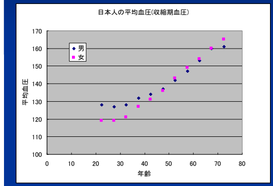
回帰と相関

イギリスの遺伝学者ガルトンの調査「人類遺伝の典型的法則」

- 親の身長とその子供の身長を調査
- 平均より背の高い人のグループ
  - 子供の平均身長は，親の平均身長ほどには高くない
- 逆に平均より低い人のグループ
  - 子供の平均身長も，親ほどには低くない
- 親の代で平均からはずれた分のうちのいくらかは子の代で戻る ⇒ 平均への回帰（regression）
- 回帰現象の効果で，人類の身長はいつの時代もほぼ一定に保たれる
- 親の身長から子供の身長を予測する式 ⇒ 回帰式

一方の変数からもう一方の変数を予測



- 30歳以上では年齢が増せば，血圧も直線的に増加の傾向
- 女性：
  血圧=140.5+1.07(年齢-50)
- 平均年齢50，
  平均血圧140.5を通る
  傾き1.07の直線

回帰直線

- $b = \dfrac{(x_1-\bar{x})(y_1-\bar{y})+\cdots+(x_n-\bar{x})(y_n-\bar{y})}{(x_1-\bar{x})^2+\cdots+(x_n-\bar{x})^2}$ 傾き
- $a = \bar{y} - b\bar{x}$ 切片
- $y = a + bx$
  $= \bar{y} + b(x-\bar{x})$ 回帰直線
- x：独立変数（説明変数）
  y：従属変数（目的変数）

内挿と外挿

- 内挿
  - 回帰直線を求めるのに使用したデータの範囲内での推測
  - 例：データはないが，45歳の血圧を推測する
- 外挿
  - 回帰直線を求めるのに使用したデータの範囲外へ直線を伸ばす
  - 例：25歳未満や75歳以上の血圧の推測
  - 範囲外の傾向が異なると適用できない

相関係数と回帰係数の検定

- 母相関係数 ρ=0 の検定
  - データ数 n と 相関係数 r から有意水準を求める
    ⇒ 表5.1 (210ページ)
  - 自由度：n-2
- 母回帰係数=0 の検定
  - 母相関係数=0 の検定と同じ計算

相関係数との関係【p.87 問2】

$$
\begin{aligned}
r &= \frac{(x_1-\bar{x})(y_1-\bar{y})+\cdots+(x_n-\bar{x})(y_n-\bar{y})}{\sqrt{(x_1-\bar{x})^2+\cdots+(x_n-\bar{x})^2}\cdot\sqrt{(y_1-\bar{y})^2+\cdots+(y_n-\bar{y})^2}} \\
&= \frac{(x_1-\bar{x})(y_1-\bar{y})+\cdots+(x_n-\bar{x})(y_n-\bar{y})}{\sqrt{n}S_x\cdot\sqrt{n}S_y} \\
&= \frac{(x_1-\bar{x})(y_1-\bar{y})+\cdots+(x_n-\bar{x})(y_n-\bar{y})}{nS_xS_y} \\
&= \frac{S_x}{S_y}\cdot\frac{(x_1-\bar{x})(y_1-\bar{y})+\cdots+(x_n-\bar{x})(y_n-\bar{y})}{nS_x^2} \\
&= \frac{S_x}{S_y}\cdot\frac{(x_1-\bar{x})(y_1-\bar{y})+\cdots+(x_n-\bar{x})(y_n-\bar{y})}{(x_1-\bar{x})^2+\cdots+(x_n-\bar{x})^2} \\
&= \frac{S_x}{S_y}\cdot b
\end{aligned}
$$

順位相関

- 相関係数は，データ(x,y)の直線関係の強さを示す尺度
- 二次関数的，指数関数的，不規則な増加（減少）をする場合
  ⇒ xの順位とyの順位の相関係数を用いる
- スペアマンの順位相関（同順位がないとき）
  - x_iの順位をR_i，y_iの順位をS_iとして

$$D=(R_1-S_1)^2+\cdots+(R_n-S_n)^2$$

$$r=1-\frac{6D}{n(n^2-1)}$$