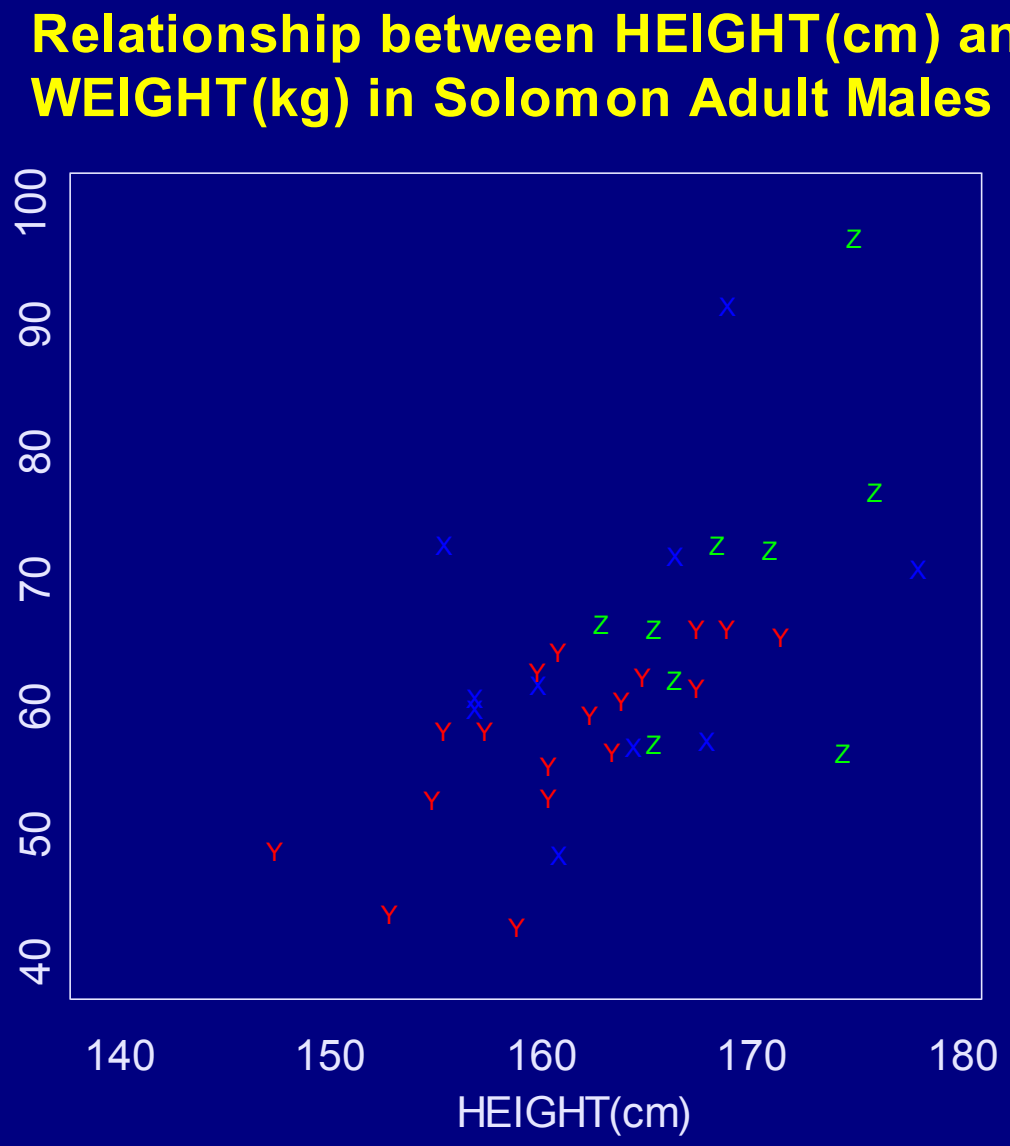
統計学第 11 回
「相関と回帰～2つの量的変数間の関係」
http://phi.ypu.jp/stat.html

- 相関は，変数間の関連の強さを表す
- 回帰は，ある変数（被説明変数，目的変数，または従属変数という）の値のばらつきが，どの程度他の変数（説明変数または独立変数という）の値のばらつきによって説明されるかを示す
- 2つの変数間の関係を予測に使うなら回帰
- 架空データ（ソロモン諸島の3つの村の男性の身長と体重の関係）で試してみる
  - タブ区切りのデータが l11-1.dat として用意してあるので，それを用いる。

まずは散布図を描く

- 2つの量的変数の関連をみるには，ともかくまず散布図（scattergram）を描いてみる
- 散布図とは，独立変数を横軸に，従属変数を縦軸にとって（たんに相関関係だけを見る場合はどちらがどちらでも良いが），二次元平面にデータ点をプロットしたもの（例：右図はRでI11-1.Rを実行した結果）。



相関の考え方

- 2個以上の変量が「かなりの程度の規則正しさをもって，増減をともにする関係」を相関関係という。
- 一方が増えると他方も増えるとき「正の相関」，一方が増えると他方が減るとき「負の相関」
- 相関関係の存在は，因果関係が存在するというための必要条件。ただし，相関があっても因果関係があるとは限らないので注意が必要。Hill の9条件は厳しすぎるので，実際に因果関係の推論をすることは難しい。疫学では，因果推論は統計によるのではなく，生物学的メカニズムによるべきであるとされる。
- 見かけの相関や擬似相関でないか注意することは大事。層別も重要。

因果関係を因果関係のない関連と区別するための Hill (1965) の基準

- 1）相関関係が強い。
- 2）相関関係が常に成り立つ。
- 3）相関関係に特異性がある。
- 4）時間的前後関係がはっきりしている。
- 5）生物学的なメカニズムが想定できる（疫学だから「生物学的」なので，主旨は社会学的でも物理的でもよい）。
- 6）もっともらしい。
- 7）首尾一貫している（他の知見と矛盾がない）。
- 8）実験的な証拠がある。
- 9）アナロジーがなりたつ。

擬似相関の例

- 日本の砂糖輸入量と溺死・溺水者数の年次別データをプロットしてみると，負の相関関係があるように見えるが，両者の間には真の関係はない。
- ある年に日本で植えた木の幹の太さと，同じ年にイギリスで生れた少年の身長を 15 年分，毎年1回測ったデータをプロットすると，おそらくは正の相関関係があるように見えるが，両者の間には関係がない。
- どちらも年次との間に真の相関があるともいえないことはないが，そもそも，このような時系列データは点と点の間が互いに独立ではないことに注意するべき（詳しくは次回）。

相関係数

- 相関の程度を示す指標が相関係数。
- 一般には直線的な（線形の）関係を示す，ピアソンの積率相関係数 r が使われる（ただ相関係数といえばこれをさす）。これは，2つの変数の共分散を，それぞれの変数の分散の積の平方根で割った値である。「相関が無い」を帰無仮説として検定するには，t 値を計算する。
- 非線形の関係については
  - 線形になるように対数変換などをしてピアソンの積率相関係数を使うか，
  - スピアマンの順位相関係数 ρ やケンドールの順位相関係数 τ などノンパラメトリックな相関係数を使う
- R では cor() で相関係数が計算できる。母相関係数がゼロという帰無仮説の検定は cor.test() でできる。

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

$$\rho = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n}(R_i - Q_i)^2$$

* Ri は Xi の，Qi は Yi の順位

$$\tau = \frac{(A-B)}{n(n-1)/2}$$

* A は順位の大小関係が一致する組数，B は不一致の組数

回帰について

- 通常は，誤差を含まない値を独立変数とし，誤差を含む測定値を従属変数とする。逆にいえば，回帰における独立変数は誤差を含まない値であると仮定されている。
- 回帰直線の推定は最小二乗法による。y=a+bx という形で推定するとき，a を切片，b を回帰係数という。
- 推定値の安定性は t 値を計算し，自由度（n−2）の t 分布を使って検定することができる。
- 従属変数のばらつきが独立変数のばらつきによって説明される割合は相関係数の二乗に一致。そこで，相関係数の二乗を決定係数とか寄与率と呼ぶ。
- 回帰直線を予測に使うときは，できるだけ外挿を避けるべき。データ点外で線形の関係が成り立つ保証はどこにもない。
- R では lm(Y~X) で回帰式が得られ，summary(lm(Y~X)) で検定結果が得られる。

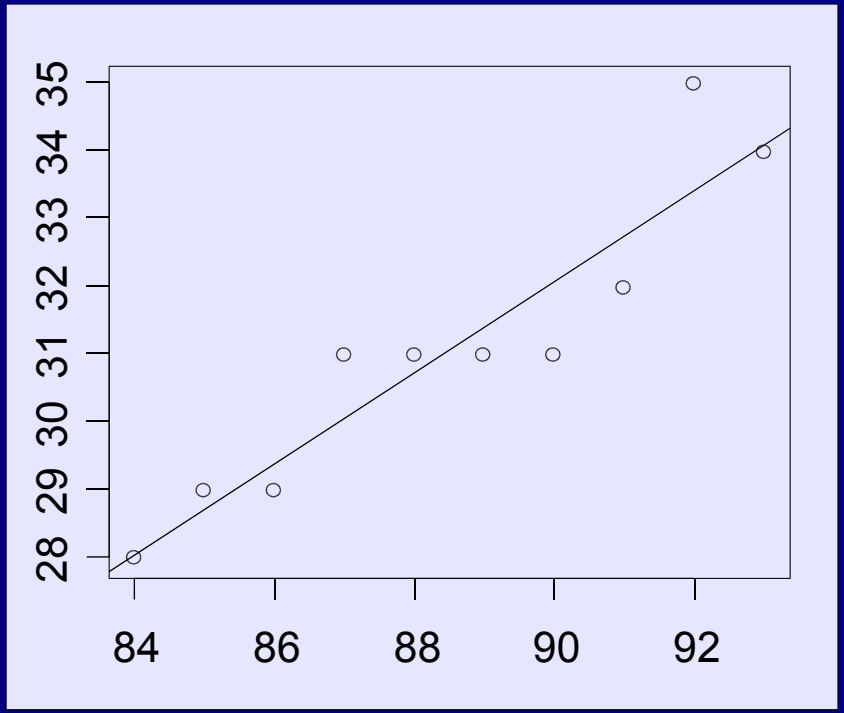
回帰の例

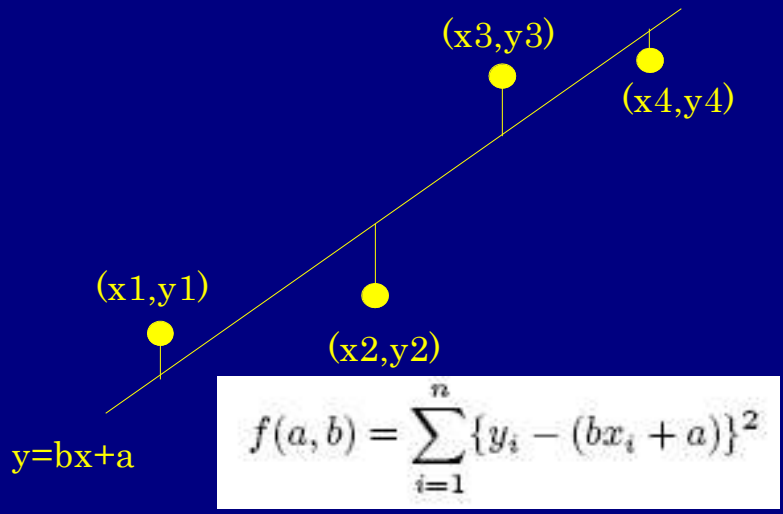
- (1) 1984 年から 1993 年までのプロ野球の1試合平均入場者数(単位:千人)の推移(出典:鈴木義一郎「情報量基準による統計解析入門」講談社)
  - プロットのプログラムは,

```
year<-84:93
c.l<-c(28,29,29,31,31,31,31,32,35,34)
plot(c.l~year)
abline(lm(c.l~year))
```

  - 年次は決まった値だが入場者数は測定値であり, 誤差を含む
- (2) 検量線:既知の濃度の標準物質を測ったときの吸光度のばらつきが, その濃度によってほぼ完全に(通常 98% 以上)説明されるときに, サンプルを測ったときの吸光度からサンプルの濃度を逆算するための回帰直線。f(a,b) を最小にする a, b を推定。

年次	84	85	86	87	88	89	90	91	92	93
セリーグ	28	29	29	31	31	31	31	32	35	34
パリーグ	13	12	16	18	21	23	22	24	24	24





回帰の外挿についての注意

- 検量線は，原則として外挿してはいけない。実際に測った濃度より濃かったり薄かったりするサンプルに対して，同じ関係が成り立つという保証はどこにもない（吸光度をｙ とする場合，濃度が高くなると分子の重なりが増えるので飽和（saturate）してしまい，吸光度の相対的な上がり方が小さくなっていき，直線から外れる）。
- 外挿による予測は，実際にはかなり行われている（ eg. 世界人口の将来予測，河川工学における基本高水計算式，感染症の発症数の将来予測，etc. ）。
- 外挿が妥当性をもつためには，次の条件が必要。
  - （1）かなり説明力が大きく，
  - （2）因果関係がある程度認められ，
  - （3）それぞれの変数の分布が端の切れた分布でない（ truncated distribution　でない）