前回のQ＆A

Q1) プリントが小さくて見にくいので元のサイズに戻してください（複数の同意見）。
A1) 紙資源節約のためと，両面印刷で５枚刷る時間が惜しかったので縮小印刷したのですが，見にくいという意見が多いようなので縮小は止めます。
Q2) 質問しようと思ったら先生はいつどこにいますか？
A2) 基本的に月曜午後から木曜夕方までは E115 にいますが，予定表を http://phi.ypu.jp/schedule.html に公開しているので参考にしてください。

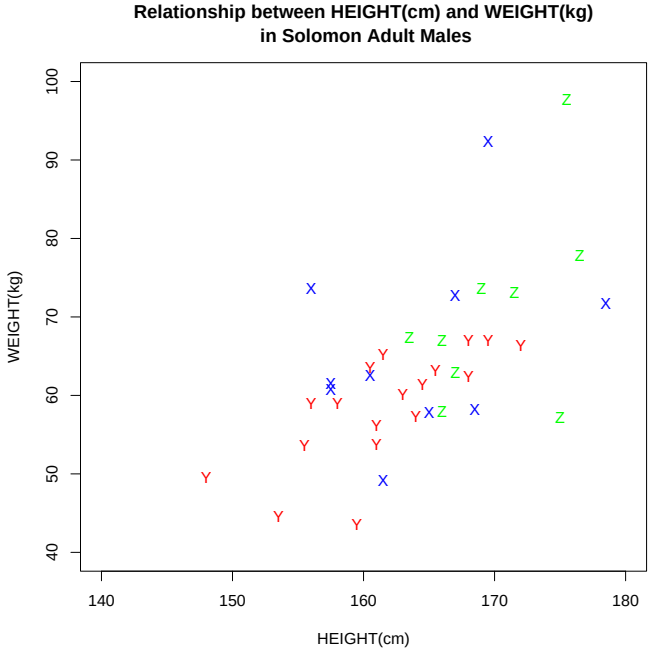
統計学第 11 回　相関と回帰

（１）量的変数の関連を調べる

- 相関と回帰は混同されやすいが，思想はまったく違う。相関は，変数間の関連の強さを表すものである。回帰は，ある変数の値のばらつきが，どの程度他の変数の値のばらつきによって説明されるかを示すものである。回帰の際に，説明される変数を従属変数または目的変数，説明するための変数を独立変数または説明変数と呼ぶ。２つの変数間の関係を予測に使うためには，回帰を用いる。
- 前回あげた南太平洋のある島にある３つの村 X, Y, Z の成人男性の身長や体重の例を使って説明する。前回は示さなかった体重のデータも加えると，データは例えば次のようになる（架空のものである）。

ID 番号	村落 (“VG”)	身長 (cm)(“HEIGHT”)	体重 (kg)(“WEIGHT”)
1	X	161.5	49.2
2	X	167.0	72.8
3	X	157.5	61.6
（中略）			
22	Z	166.0	58.0
（中略）			
36	Y	156.0	59.0
37	Y	155.5	53.6

- 身長と体重の関係を，身長を横軸にとって，体重を縦軸にとって二次元平面にプロットすると，下図のようになる[1] 。第３回の講義で触れたが，このような図を散布図 (scatter plot または scattergram) と呼ぶ。２つの量的変数間の関係をみるときには，基本として絶対に作成しなければならない。



- 関係とか関連とかいっても，その中身は多様である。例えば，$pV = nRT$ のような物理法則は，測定誤差を別にすれば 100%成り立つ関係である。身長と体重の間の関係はそうではないが，無関係

[1] R を使って，村ごとにプロットするマークを変えてプロットした。やり方に関心があれば，l11-1.R をダウンロードして参照されたい。

ではないことは直感的にも理解できるし，上の図を見ても「身長の高い人は体重も概して重い傾向がある」ことは間違いない。

- 一般に，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」のことを相関関係 (correlation) という。相関には正の相関 (positive correlation) と負の相関 (negative correlation) があり，一方が増えれば他方も増える場合を正の相関，一方が増えると他方は減る場合を負の相関と呼ぶ。例えば，上の図に示されている身長と体重の関係は正の相関である。
- 相関関係があることは，因果関係 (causal relationship) が成り立つための重要な要件ではあるが，それだけで因果関係があるとは結論付けるのは勇み足である。では，因果関係があるというための基準はあるのだろうか？　古来いろいろな説があった中でも有力とされていて，疫学の標準的な教科書である Rothman の”Modern Epidemiology”にも載っている Hill(1965) の基準によれば，因果関係を因果関係のない関連と区別するためには，次の９条件が満たされる必要がある。[2]
  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（これは疫学の教科書での説明だから「生物学的な」なので，社会学的であっても分子的であってもよい）。
  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。
- 相関関係があっても，それが見かけ上のものである（それらの変量がともに，別の変量と真の相関関係をもっている）場合がある。具体例としては，血圧と所得の間に正の相関があるという命題は，データをとってみれば，多くの場合に成り立つであろう。しかしこれは，おそらくどちらも年齢や摂取エネルギー量との間に真の相関関係が存在するのであって，それらの影響を制御したら（例えば同年齢で同じような食生活をしている人だけについて見る，という層別化をしたら），血圧と所得の間の正の相関は消えてしまうだろう。この場合，見かけ上の相関があることは，たまたまそのデータで成り立っているだけであって，科学的仮説としての意味に乏しい。因果関係に迫ることが大事なのであって，相関関係はその入り口に過ぎないことを再び強調しておく。
- 時系列データや地域相関のデータでは，擬似相関 (spurious correalation) が見られる場合もある。例えば，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のだが，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のだが，両者の間には関係がないのは明らかである（どちらも年次と真の相関があるともいえるが）。
- 複数の種類の異なるデータをまとめて見ることで見かけの相関が生じてしまう場合もある。上に示した南太平洋の３つの村の身長と体重の関係を良く見ると，相関関係は村によって随分違っていることがわかる。それをまとめたことで身長の分散と体重の分散が広がって，見かけ上強い正の相関がでたと解される。こういう場合は，村で層別して村ごとに相関を検討する必要がある[3] 。

（２）相関関係の具体的な捉え方

- 上で定義したように，相関関係は増減をともにする関係であればいいので，その関係が線形（一次式で表される，散布図で直線として表される）であろうと非線形（二次式以上または階段関数などで表される）であろうと問題ない。しかし，一般には，線形の関係があるという限定的な意味で使われる場合が多い。なぜなら，相関を表すための代表的な指標である相関係数（普通，ただ相関係数といえば，ピアソンの積率相関係数 (Pearson’s Product Moment Correlation Coefficient) を指す）

[2] もっとも，Hill 自身が，必ずしも９条件すべてが成り立たない場合もあるし，ヒトについては実験的な証拠が得られない場合が普通であるなど，この条件が決定的とすることは因果推論の現実性を失わせてしまうことも認めているので，あくまでこうした基準は目安程度に考えるべきである。なかでも，相関関係の特異性という考え方は，因果推論の適用範囲を著しく狭めてしまう。現実の因果関係の大部分は，複数の要因と複数の結果が網の目のように絡んでいるし，環境によって同じ遺伝的要因がまったく逆の結果をもたらす場合もある。だからこそ Rothman は Modern Epidemiology や，2002 年に発表した入門書である”Epidemiology: An Introduction”（因果推論について書かれた第２章が無料で全文，`http://www.oup-usa.org/sc/0195135547/0195135547_ch2.pdf` としてダウンロードできる）において，簡単な基準は存在しないことを強調し，因果構成要因群 (Component causes) と充分要因 (Sufficient cause) という考え方（第１回の講義を参照のこと）を提起したのである。

[3] または，目的次第では，ダミー変数を使った重回帰分析をするべきである（第 13 回の講義で触れる予定）。

r が，線形の関係を示すための指標だからである。もっといえば，r が意味をもつためには，２つの変量が二次元正規分布に従っていなければならない。

- 非線形の相関関係を捉えるには，２つのアプローチがある。１つは線形になるように対数変換などの変換をほどこすことで，もう１つはノンパラメトリックな相関係数（順位の情報だけを使った相関係数）を使うことである。ノンパラメトリックな相関係数にはスピアマン (Spearman) の順位相関係数 ρ や，ケンドール (Kendall) の順位相関係数 τ がある。
- ピアソンの積率相関係数とは，X と Y の共分散を X の分散と Y の分散の積の平方根で割った値である。式で書けば，相関係数の推定値 r は，X の平均を $\bar{X}$，Y の平均を $\bar{Y}$ と書けば，

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

となる。母相関係数がゼロかどうかという両側検定のためには，それがゼロであるという帰無仮説の下で，検定統計量

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

が，自由度 $n-2$ の t 分布に従うことを利用して検定すればよい。

- なおＲでは，`r _ cov(X,Y)/sqrt(var(X)*var(Y)); n _ NROW(X); t0 _ r*sqrt(n-2)/sqrt(1-r ^ 2);` として，`2*(1-pt(t0,n-2))` で有意確率が得られるが，`cor.test` 関数（下記）を使う方が簡単である。`cor.test` 関数を使った場合は，信頼区間も計算される。なお，信頼区間は，サンプルサイズがある程度大きければ（通常は 20 以上），正規近似を使って計算できる。

$$a = \frac{1}{2}\ln\frac{1+r}{1-r} - \frac{1}{\sqrt{n-3}}Z(\alpha/2), \quad b = \frac{1}{2}\ln\frac{1+r}{1-r} + \frac{1}{\sqrt{n-3}}Z(\alpha/2)$$

と書くことにすると（$Z(\alpha/2)$ は，標準正規分布の $100 \times (1-\alpha/2)$ パーセント点，つまり，Ｒでは `qnorm(1 - α/2,0,1)` である），母相関係数の $100 \times (1-\alpha)$%信頼区間の下限は $(\exp(2a)-1)/(\exp(2a)+1)$，上限は $(\exp(2b)-1)/(\exp(2b)+1)$ である[4] 。

- 順位相関係数は，非線形の相関関係を捉えたい場合以外にも，分布が歪んでいたり，外れ値がある場合に使うと有効である。スピアマンの順位相関係数 ρ は[5] ，値を順位で置き換えた（同順位には平均順位を与えた）ピアソンの積率相関係数になる。X_i の順位を R_i，Y_i の順位を Q_i とかけば，

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

となる。スピアマンの順位相関係数がゼロかどうかという両側検定は，サンプル数が 10 以上ならばピアソンの場合と同様に，$T = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}}$ が自由度 $n-2$ の t 分布に従うことを利用して行うことができる。

- ケンドールの順位相関係数 τ は，

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

によって得られる。ここで A は順位の大小関係が一致する組の数，B は不一致数である。

- いずれにせよ，Ｒ では `cor.test(X, Y, method="pearson")` とすればピアソンの相関係数が，`cor.test(X, Y, method="spearman")` でスピアマンの順位相関係数が，`cor.test(X, Y, method="kendall")` でケンドールの順位相関係数が得られる。同時に，`alternative` を指定しないときは,「相関係数がゼロである」を帰無仮説として両側検定した有意確率と 95%信頼区間が表示される。なお，例えば `cor.test(X, Y, alternative="g")` とすれば，ピアソンの相関係数が計算され，対立仮説を「正の相関がある」とした片側検定の結果が得られる。なお，ケンドールに関しては並べ換えによる正確な確率も求めることができ，その場合は `exact=T` というオプションを指定する。

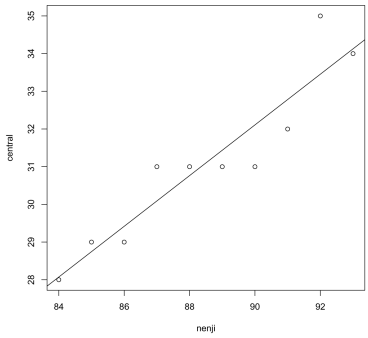
[4] なお，ln は自然対数，exp は指数関数を表す。

[5] ピアソンの相関係数の母相関係数を ρ と書き，スピアマンの順位相関係数を r_s と書く流儀もある。

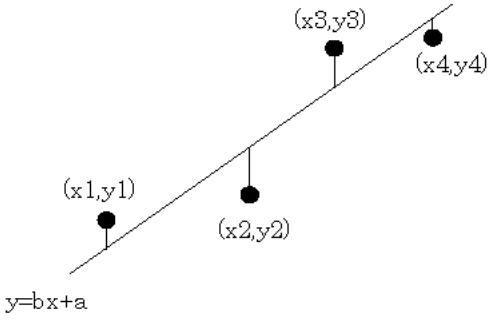
（3）回帰の考え方

- 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）の推移は下表のようになっている（注：この表の出典は，鈴木義一郎「情報量基準による統計解析入門」講談社サイエンティフィク，である）。

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



- これを見ると，途中から伸び悩んでいるが，ある程度直線的に増加しているように見える。このように，ほぼ比例関係にある量的なデータ群をうまく代表する直線を求めるには,「各データの点から直線までのずれの大きさの合計」を最小にすればよい。この場合,「年次」が予め決まっている値であるのに対して，入場者数は測定値であり，誤差を含む可能性があるので,「データ点から直線までのずれ」を評価するには，データ点と直線の最短距離よりも，データ点から直線に垂直に下ろした線分の長さの二乗和を使う方がよい。この，ずれを最小にする直線を,「年次を独立変数，入場者数を従属変数とする回帰直線」と呼ぶ。計算方法は，数学的には下記の検量線の求め方と同じである。R では，`nenji ← c(84:93); central ← c(28,29,29,31,31,31,31,32,35,34); plot(nenji,central); z ← lm(central ~ nenji); abline(z)` とすれば上の図が描かれる。
- 実験によって，あるサンプルの濃度を求めるやり方の 1 つに，検量線の利用がある。検量線とは，予め濃度がわかっている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，その回帰を利用して，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である（曲線の場合もあるが，通常は何らかの変換をほどこし，線形回帰にして利用する）。検量線の計算には，(A) 試薬ブランクでゼロ点調整をした場合の原点を通る回帰直線を用いる場合と，(B) 純水でゼロ点調整をした場合の切片のある回帰直線を用いる場合がある。いずれも，量がわかっているもの（この場合は濃度）を x，誤差を含んでいる可能性がある測定値（この場合は吸光度）を y として $y = bx + a$ という形の回帰式を最小二乗法で推定し，サンプルを測定した値 y から $x = (y - a)/b$ によってサンプルの濃度 x を求める。回帰直線の適合度の目安としては，学生実習でも相関係数の 2 乗が 0.98 以上あることが望ましい。



- 図のような測定点 (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) が得られたときに，検量線 $y = bx + a$ を推定するには，図に示した線分の二乗和が最小になるように a と b を設定すればよい，というのが最小二乗法の考え方である。つまり，

$$f(a,b) = \sum_{i=1}^{n} \{y_i - (bx_i + a)\}^2$$

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

となるような a と b を推定すればよい。通常，a と b で偏微分した値がそれぞれ 0 となることを利用して計算すると簡単である。つまり，

$$\frac{\partial f(a,b)}{\partial a} = 2na + 2(b \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i) = 0$$

$$i.e. \quad na = \sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} x_i$$

$$\frac{\partial f(a,b)}{\partial b} = 2b \sum_{i=1}^{n} x_i^2 + 2(a \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i) = 0$$

$$i.e. \quad b \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i - a \sum_{i=1}^{n} x_i$$

を連立方程式として a と b について解けばよい。これを解くと，

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

が得られ，これを上の式に代入すれば a も得られる。検量線に限らず，一般の回帰直線でも，計算方法は原則として同じである。名称の説明をしておくと，一般に，$y = bx + a$ という回帰直線について，b を回帰係数 (regression coefficient)，a を切片 (intercept) と呼ぶ。
- データから得た回帰直線は，$pV = nRT$ のような物理法則と違って，完璧にデータに乗ることはない。そこで，回帰直線の当てはまりのよさを評価する必要が出てくる。いま，$z_i = a + bx_i$ とおいたときに，$e_i = y_i - z_i$ を残差 (residual) と呼ぶ。残差は，y_i のばらつきのうち，回帰直線では説明できなかった残りに該当する。つまり，残差が大きいほど，回帰直線の当てはまりは悪いと考えられる。残差にはプラスもマイナスもあるので，例によって二乗和をとって，

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{n=1}^{n} (y_i - z_i)^2 = \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2/n - \frac{(n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} xi \sum_{i=1}^{n} y_i)^2}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}/n$$

が，回帰直線の当てはまりの悪さを示す尺度となる。この Q を「残差平方和」と呼び，それを n で割った Q/n を残差分散という。この残差分散 $var(e)$ と Y の分散 $var(Y)$ とピアソンの相関係数 r の間には，$var(e) = var(Y)(1 - r^2)$ という関係が常に成り立つので，$r^2 = 1 - var(e)/var(Y)$ となる。このことから r^2 が 1 に近いほど回帰直線の当てはまりがよいことになる。その意味で，r^2 を「決定係数」と呼ぶ。また，決定係数は，Y のばらつきがどの程度 X のばらつきによって説明されるかを意味するので，X の「寄与率」と呼ぶこともある。
- 回帰直線は最小二乗法でもっとも残差平方和が小さくなるように選ぶわけだが，データの配置によっては，何通りもの回帰直線の残差平方和が大差ないという状況がありうる。例えば，独立変数と従属変数（として選んだ変数）が実はまったく無関係であった場合は，データの重心を通るどのような傾きの線を引いても残差平方和はほとんど同じになってしまう。その意味で，回帰直線のパラメータ（回帰係数 b と切片 a）の推定値の安定性を評価することが大事である。そのためには，t 値というものが使われている。いま，Y と X の関係が $Y = a_0 + b_0 X + e$ というモデルで表されるとして，誤差項 e が平均 0，分散 σ^2 の正規分布に従うものとすれば，回帰係数の推定値 a も，平均 a_0，分散 $\sigma^2/n)(1 + M^2/V)$（ただし M と V は x の平均と分散）の正規分布に従い，残差平方和 Q を誤差分散 σ^2 で割った Q/σ^2 が自由度 $(n-2)$ のカイ二乗分布に従うことから，

$$t_0(a_0) = \frac{\sqrt{n(n-2)}(a - a_0)}{\sqrt{(1 + M^2/V)Q}}$$

が自由度 $(n-2)$ の t 分布に従うことになる。しかしこの値は a_0 がわからないと計算できない。a_0 が 0 に近ければこの式で $a_0=0$ と置いた値（つまり $t_0(0)$。これを切片に関する t 値と呼ぶ）を観測データから計算した値が $t_0(a_0$ とほぼ一致し，自由度 $(n-2)$ の t 分布に従うはずなので，その絶対値は 95%の確率で t 分布の 97.5%点（サンプルサイズが大きければ約 2 である）よりも小さくなる。つまり，データから計算された t 値がそれより大きければ，切片は 0 でない可能性が高いことになる。言い換えると,「切片が 0 である」という帰無仮説に対する有意確率が計算できることになる。回帰係数についても同様に,

$$t_0(b)=\frac{\sqrt{n(n-2)V}b}{\sqrt{Q}}$$

が自由度 $(n-2)$ の t 分布に従うことを利用して,「回帰係数が 0」であるという帰無仮説に対する有意確率が計算できる。

- 以上の説明からすると，身長と体重のように，どちらも誤差を含んでいる可能性がある測定値である場合には，どちらかを独立変数，どちらかを従属変数とみなしてよいのかということが問題になってくる。一般には，身長によって体重が決まってくるというように方向性が仮定できれば，身長を独立変数と見なしてもよいことになっているが，回帰分析をしてしまうと，独立変数に測定誤差がある可能性が排除されてしまうことには注意しておくべきである。つまり，測定誤差が大きい可能性がある変数を独立変数とした回帰分析は，できれば避けたほうが良い（が，そうもいかないのが実情である）。また，最小二乗推定の説明から自明なように，独立変数と従属変数を入れ替えた回帰直線は一致しないので，どちらを従属変数とみなし，どちらを独立変数とみなすか，ということは，因果関係の方向性に基づいてきちんと決めるべきである。
- ところで，回帰 (regression) とは（１）で説明した通り，被説明変数（従属変数）のばらつきが，説明変数（独立変数）のばらつきで説明されるという考え方だが，もともとは，生物統計学者 Francis Galton が，父親と息子の身長をペアとして測定し，背の高い父親の息子の平均身長が父親ほど高くなく，背の低い父親の息子の平均身長が父親ほど低くないこと，つまり第二世代の身長が平均の方向に「回帰」するという意味で用いた言葉である。この現象は，父親群でも息子群でも身長の平均と分散が等しいと仮定し，父親の身長と息子の身長の分布が二次元正規分布に従うとすると以下のようにクリアに説明できる。
- 父親の身長が x の息子の身長 Y の期待値 $\mu_{Y\cdot x}$ は，父親の身長と息子の身長の母相関係数を R と書くことにすると，$\mu_{Y\cdot x}=\mu_y+R\frac{\sigma_y}{\sigma_x}(x-\mu_x)$ となるので，これを式変形すれば，$\mu_{Y\cdot x}-x=-(1-R)(x-\mu_x)$ となるので，$x>\mu_x$ ならば $\mu_{Y\cdot x}<x$ となり，$x<\mu_x$ ならば $\mu_{Y\cdot x}>x$ となる。この式は，Galton が観察した現象と符合する。
- 回帰を使って予測をするとき，外挿には注意が必要である。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からである（吸光度を y とする場合は，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，直線から外れていく）。しかし，外挿による予測は，実際にはかなり行われている。例えば世界人口の将来予測とか，河川工学における基本高水計算式とか，感染症の発症数の将来予測は，回帰の外挿による場合が多い。このやり方が妥当性をもつためには，その回帰関係が（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）という条件を満たす必要がある。そうでない場合は，その予測結果が正しい保証はどこにもない[6] 。
- R では，今回説明したような線形回帰を行うための関数は `lm` である。例えば，`lm(Y ~ X)` のように用いれば，回帰直線の推定値が得られる。決定係数や回帰係数と切片の検定結果は，`summary(lm(Y ~ X))` とすれば出力される（他の統計ソフトでも簡単に得られるはずである）。なお，普通の量的変数の間の線型回帰を一般化すれば，t 検定，分散分析，共分散分析，回帰分析，判別分析，正準相関分析などをすべて共通の数学モデルで扱うことができる。このモデルを一般化線型モデルと呼ぶ。英語では General Linear Model といい，R での関数名も `glm` である。一般化線形モデルは，基本的には，$Y=\beta_0+\beta X+\varepsilon$ という形で表される（Y が従属変数群，X が独立変数群，β_0 が切片群，β が係数群，ε が誤差項である）。

[6] それでも簡便さのために回帰の外挿による予測はかなり行われてしまっているのが現状だが，本来そういう場合は，単純な回帰でなく確率的な因果モデルを立て，シミュレーションを行うべきである