2026 年度春学期　統計学　第 7 回

データの関係を知る (2) ー 回帰と決定係数

回帰分析

回帰分析は，２つ以上の変量の組で表されるデータがあるとき，ある変量と他の変量との関係を求める方法のひとつです。「関連の強さ」を調べる相関分析と違い，回帰分析では，一方の変量によって他方の変量が決まるという関係があるとき，「ある変量の変化を，もう一方の変量の変化で説明するための関数を求める」という考え方をします。今回は，もっとも基本的な回帰分析である線形単回帰について説明します。

線形単回帰ー直線のあてはめ

回帰分析と線形単回帰

前回の講義で用いた，各都市の緯度と気温のデータ，およびその散布図をもう一度見てみましょう（表 1，図 1)。散布図上のデータは，好き勝手にばらついているわけではありません。前回説明したように，緯度と気温の間には負の相関関係があります。そこで，これらのデータのばらつき方を，**気温が緯度から計算で求められるというモデル**で表現しようというのが**回帰分析**です。

緯度を x とし，気温を y とするとき，「x から y が計算で求められる」という関係になっていることを統計学では「y は x によって説明される」といい，x を説明変数，y を被説明変数といいます。また，この関係を y の x 上への**回帰**といいます。この例の場合，明らかに散布図上で右下がりの直線となるような関係がありそうです。だからといって，散布図上に＋印の列が完全に直線上に並んでいるわけでもありません。では，どういう直線をひけばよいのでしょうか。

緯度 x と気温 y に散布図上で直線の関係があると仮定するということは，散布図上にばらついているデータを，$y = a + bx$ という式（つまり計算）で表される直線というモデル，すなわち線形モデルで表すことになります。このような回帰を，**線形単回帰**といいます。

回帰直線の式を求めるには

そこで，この式の a, b つまりパラメータを決める方法を考えます。与えられている緯度と気温の組を (x_i, y_i) とします。x と y の間の関係が，$y = a + bx$ というモデルで完全に説明されるのなら，$x = x_i$ のとき $y = a + bx_i$ となるはずです。しかし，現実には $y = y_i$ となっています。そこで，パラメータのさまざまな値のうちで，この「全ての (x_i, y_i) についての，y_i と $a + bx_i$ との差の合計」が，もっとも小さくなるパラメータをもっとも適切なパラメータとします。差には正負がありますから，実際には差の２乗の合計，すなわち

$$L = \sum_{i=1}^{n} \{y_i - (a + bx_i)\}^2 \tag{1}$$

が最小になるように a と b を決定します（n はデータの組の数です)。

表 1: 日本の都市の緯度と気温

地名	緯度（度）	気温（℃）
札幌	43.05	8.0
青森	40.82	9.6
秋田	39.72	11.0
仙台	38.27	11.9
福島	37.75	12.5
宇都宮	36.55	12.9
水戸	36.38	13.2
東京	35.68	15.3
新潟	37.92	13.1
長野	36.67	11.4
静岡	34.97	16.0
名古屋	35.17	14.9
大阪	34.68	16.2
鳥取	35.48	14.4
広島	34.40	15.0
高知	33.55	16.3
福岡	33.92	16.0
鹿児島	31.57	17.3
那覇	26.20	22.0

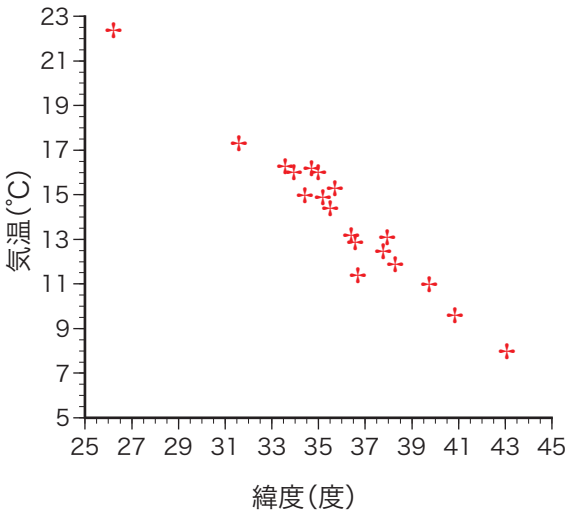


図 1: 散布図：緯度と気温の関係

「偏微分」で求める

このような a と b を求める方法は，おもに 2 つあります。ひとつは，(1) 式を a と b でそれぞれ「偏微分」し，それらを両方とも 0 とおいた方程式を解くものです。

「a と b それぞれで偏微分する」とは，次のような意味です。微分とは，関数のグラフ上のある点での接線の傾きを求めることです。そこで，(1) 式の L を a, b の 2 つの変数の関数と考えると，この関数は a, b のどちらについても 2 次関数で，a^2, b^2 の係数がいずれも正ですから，そのグラフは a, b どちらの軸でみても下に凸の放物線で，すなわち図 2 のような曲面になります。「a と b それぞれで偏微分する」というのは，L を a だけの関数・b だけの関数とみなしてそれぞれ微分することで，曲面上のある点で，a 軸方向の接線の傾き・b 軸方向の接線の傾きを求めることになります。曲面上で，どちらの偏微分も 0 になる点は，曲面の底にしかありません。ですから，どちらの偏微分も 0 になるときの a, b の値が，L を最小にする a, b の値です。この方法で a, b を求める方法は，付録 1 で説明しています。

偏微分を使うには，大学の理科系学部の 1 年生で習うくらいの解析学の知識が必要です。(1) 式が a, b のそれぞれの 2 次関数であることから，「2 次関数の最大・最小」を使えば，この問題を高校数学の知識で解くこともできます。この方法は，付録 2 で説明しています。

回帰係数，回帰直線

いずれの方法でも，結果は

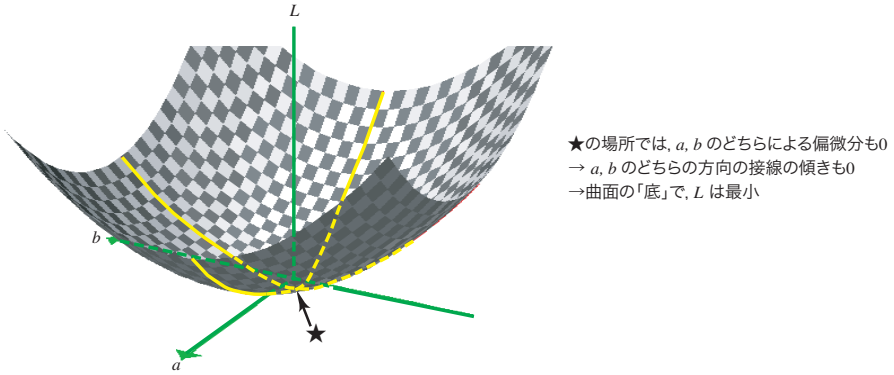


図 2: 偏微分と関数の最小値

$$\begin{aligned} b &= \frac{\sigma_{xy}}{\sigma_x^2} \\ a &= \bar{y} - b\bar{x} \end{aligned} \tag{2}$$

となります。ここで，σ_x^2 は x の分散，σ_{xy} は x, y の共分散です。$\bar{x}, \bar{y}$ は，前回も出てきたもので，それぞれ x の平均，y の平均です。

この方法を**最小二乗法**といい，このようにして得られる 1 次式 $y = a + bx$ を y の x 上への**回帰方程式**，あるいは**回帰直線**といいます。また，b は回帰直線の傾きで，これを**回帰係数**といいます。なお，(2) 式を $y = a + bx$ に代入すると

$$y - \bar{y} = b(x - \bar{x}) \tag{3}$$

となりますから，散布図上で回帰直線は「傾きが b で点 $(\bar{x}, \bar{y})$ を通る直線」になります。

「緯度と気温」の例で

さきほど表 1 であげた緯度と気温の例を考えます。緯度を変量 x, 気温を変量 y として回帰直線 $y = a + bx$ を求めると，$b = -0.850, a = 44.60$ となり，たしかに図 1 の散布図に並んだ点に沿った，右肩下がりの直線になります。

また，この結果を使って，緯度から気温を計算で推定してみましょう。緯度が 35.0 度の都市があるとすると，気温は $y = a + bx$ に $x = 35.0$ を代入して $44.60 + (-0.850) + 35.0 = 14.85(℃)$ と推定されます。表の中にある都市を見ると，大阪市は緯度が 34.68 度で，気温は $44.60 + (-0.850) + 34.68 = 15.12(℃)$ と推定されます。実際の気温は 16.2 ℃で，推定された気温よりも高くなっています。

残差と決定係数

残差

各 x_i に対して，回帰直線上で対応する y の値，すなわち $a+bx_i$ を $\hat{y}_i = a+bx_i$ と表すことにします。このとき，実際のデータにおける y_i と $\hat{y}_i$ の差を**残差**といい，d_i で表します。残差は，回帰方程式と x_i の値を使って y_i の値を $\hat{y}_i$ と推定したとき，推定によって表現できなかった部分を表しています。前節の最後で推定した大阪市の気温でいうと，大阪市が i 番の都市だとすれば，推定された気温 $\hat{y}_i$ が 15.12 ℃で，実際の気温 y_i は 16.2 ℃ですから，残差は $16.2 - 15.12 = 1.08$(℃) となります。

決定係数

残差について，r_{xy} を x と y の相関係数（前回の講義参照）とすると

$$\sum d_i^2 = \sum (y_i - \hat{y}_i)^2 = (1 - r_{xy}^2) \sum (y_i - \bar{y})^2 \tag{4}$$

が成り立ちます（導出は付録３）。さらにこの式を変形して

$$1 - r_{xy}^2 = \frac{\sum d_i^2}{\sum (y_i - \bar{y})^2} = \frac{\sum d_i^2/n}{\sum (y_i - \bar{y})^2/n} \tag{5}$$

としてみます。相関係数の 2 乗である r_{xy}^2 を，**決定係数**といいます。決定係数には，どういう意味があるのでしょうか？

決定係数の意味

(5) 式の右端の分母は，y 全体の平均からの各 y_i のへだたり，すなわち偏差の２乗の平均で，つまり y の分散を表しています。一方，分子は，残差の２乗の平均になっています。残差は「線形モデルによる予測結果からの隔たり」ですから，分子は「線形モデルによる予測結果を中心とするばらつき具合」を表しています（図 3)。

したがって，$(1 - r_{xy}^2)$ は「もともとの y のばらつき具合に対する，線形モデルからのばらつき具合の割合」を示す値ということになります。線形単回帰では，「データが散布図上にばらついている」という状況を，「好き勝手にばらついているのではなく，線形モデルで表される直線に沿ってばらついている」と説明しています。しかし，線形モデルで完全に表されたわけではなく，直線から見てもデータはいくらかばらついていますから，上の説明で完全に説明がついているわけではありません。こう考えると，r_{xy}^2 は「直線からのばらつきは，もともとあった y の分散に比べて，何%減少しているのか」を示す値ですから，r_{xy}^2 は「線形単回帰によって，データのばらつきの何%の説明がついたか」を表しています。

もし $r_{xy}^2 = 1$ ならば，分散が 100%減少して残差 $= 0$ ということですから，データのばらつきは線形単回帰によって 100%説明がついた，ということを意味しています。これは，相関係数 $= \pm 1$ のときに，散布図上の点が直線上に完全に並んでいることに対応しています（図 4)。

「中程度の相関」とは

前回の講義で，「相関係数 $r_{xy} = 0.5$ は，中程度の相関ではなくほとんど相関が無いことを示す。相関係数 $r_{xy} = 0.7$ であれば，一応相関があるといえる」という説明をしましたが，その根拠はこの決定係数にあります。相関係数 $r_{xy} = 0.5$ のとき，決定係数 $r_{xy}^2 = 0.25$ ですから 25% の減少で，もとの y の分

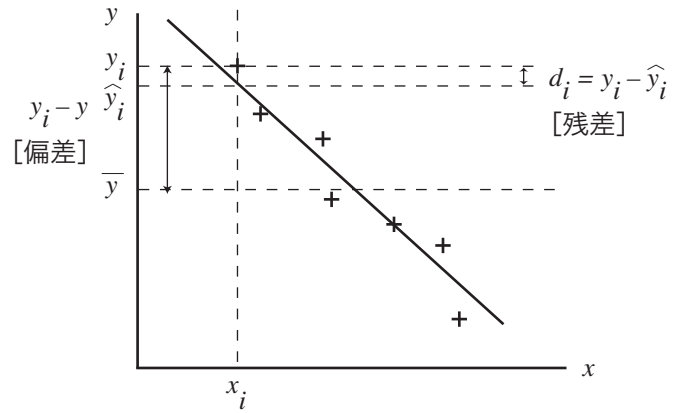


図 3: 偏差と残差

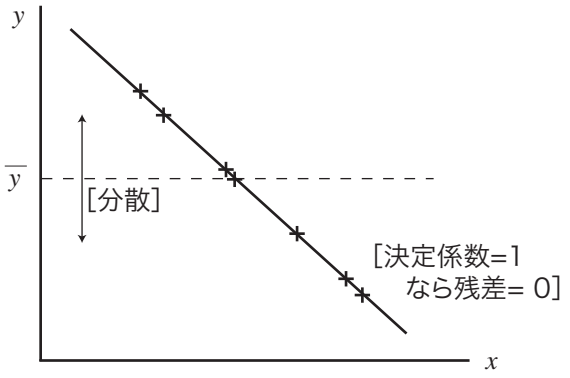


図 4: 決定係数の意味

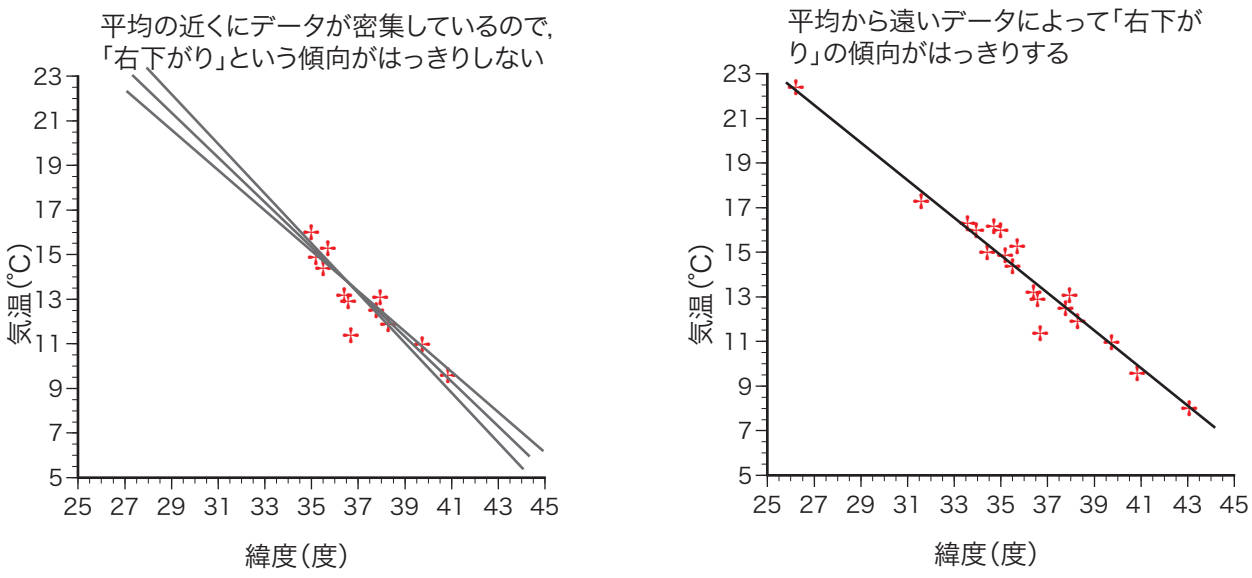
散の 75%は回帰直線からの残差にそのまま残っています。相関係数が 0.7 以上であれば，決定係数はほぼ 0.5 以上になって，回帰直線からのばらつきはもとの分散の半分以下になるので，確かに回帰直線を引く意味がある，すなわち，線形モデルで表すことに意味があるほどの，はっきりとした相関があるといえることになります。

データ圧縮の基盤へ

ところで，y が x によって完全に正確に決定される，つまり決定係数が 1 であるということは，言い方を変えれば「(x_i, y_i) の組になっているデータのうち，x_i さえわかれば，y_i は計算で求められるから，データとして記録する必要がない」ことを意味します。また，決定係数が 1 に近ければ，「x_i がわかれば，y_i の値はほぼ見当がつく」ことになります。このような考え方は，情報科学におけるデータ圧縮の基盤となっています。

データの一部を使ったときの決定係数の違い

前回・今回の講義で用いている，都市の緯度と気温のデータに関して，長野〜鹿児島の数値だけを使って相関係数を計算すると -0.844 で，札幌〜那覇のデータ全体を使って相関係数を計算すると -0.974 となります。したがって，前者の決定係数は $(-0.844)^2 = 0.712$，後者の決定係数は $(-0.974)^2 = 0.949$ となります。データの大半は共通なのにこのような違いがあるのは，後者のほうには，札幌・那覇という平均から離れたデータがあるために，「散布図のうえで右下がり」という傾向がよりはっきりしていることによります（図 5)。



(a) 長野〜鹿児島のデータを使った場合　　(b) 札幌〜那覇のデータを使った場合

図 5: データの一部を使ったときの決定係数の違い

「直線からのばらつきは，もともとあった y の分散に比べて，何%減少しているのか」ということの意味を，もう少し説明してもらえませんか？

回帰分析というのは，緯度と気温の「関係」を，この場合は線形モデルで説明する，という方法やったな。そうすると，直線からのばらつきというのは，線形モデルでは説明しきれなかった部分ということになる。「説明しきれなかった部分」がもとのデータのばらつきに比べて大きく減少しているほうが，回帰分析は「成功」というわけで，決定係数は回帰分析の成功の度合，というわけやな。

演習問題

1. 決定係数の意味を，「分散を説明する」という言葉を使って説明してください。
2. X, Y の 2 つの変量からなるデータがあり，変量 X から変量 Y への回帰直線を求めるとします。このとき，X と Y の間に相関がないとすると，回帰直線は散布図上でどのような直線になるでしょうか。式を使って理由を説明して答えてください。

付録

※以下の「付録」は，試験の範囲には入りません。

付録１：偏微分を用いた最小二乗法の導出

(1) 式を展開すると（以下，$\sum$ の添字を省略します），

$$L = \sum_{i=1}^{n} \{y_i - (a + bx_i)\}^2 = \sum y_i^2 - 2b\sum x_i y_i - 2a\sum y_i + na^2 + 2ab\sum x_i + b^2 \sum x_i^2 \qquad \text{(A1)}$$

であり，L を a, b でそれぞれ偏微分した $\dfrac{\partial L}{\partial a}, \dfrac{\partial L}{\partial b}$ をそれぞれ 0 とおくと

$$\begin{aligned} \frac{\partial L}{\partial a} &= -2\sum y_i + 2na + 2b\sum x_i = 0 \\ \frac{\partial L}{\partial b} &= -2\sum x_i y_i + 2a\sum x_i + 2b\sum x_i^2 = 0 \end{aligned} \qquad \text{(A2)}$$

となり，それぞれ整理すると，

$$\begin{aligned} na + (\sum x_i)b &= \sum y_i \\ (\sum x_i)a + (\sum x_i^2)b &= \sum x_i y_i \end{aligned} \qquad \text{(A3)}$$

という連立方程式（正規方程式といいます）が得られます。ここで，x, y それぞれの平均を

$$\bar{x} = \frac{\sum x_i}{n},\ \bar{y} = \frac{\sum y_i}{n} \qquad \text{(A4)}$$

とおいて代入すると

$$\begin{aligned} na + n\bar{x}b &= n\bar{y} \\ n\bar{x}a + (\sum x_i^2)b &= \sum x_i y_i \end{aligned} \qquad \text{(A5)}$$

となります。(A5) 式の上段の式から

$$a = \bar{y} - b\bar{x} \qquad \text{(A6)}$$

が得られます。また，(A5) 式の上段の式を $\bar{x}$ 倍して下段の式から引くと

$$(\sum x_i^2 - n\bar{x}^2)b = \sum x_i y_i - n\bar{x}\bar{y} \tag{A7}$$

となるので,

$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} \tag{A8}$$

が得られます。

ところで，x の分散を σ_x^2 とすると，

$$\begin{aligned} \sigma_x^2 &= \frac{1}{n}\sum (x_i - \bar{x})^2 \\ &= \frac{\sum x_i^2}{n} - 2\bar{x}\frac{\sum x_i}{n} + \bar{x}^2 \\ &= \bar{x^2} - \bar{x}^2 \end{aligned} \tag{A9}$$

ですから，(A8) 式の分母は $n(\bar{x^2} - \bar{x}^2) = n\sigma_x^2$ となります。また，x, y の共分散を σ_{xy} とすると

$$\begin{aligned} \sigma_{xy} &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n} \\ &= \frac{1}{n}(\sum x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y}) \\ &= \frac{\sum x_i y_i}{n} - \bar{x}\bar{y} \end{aligned} \tag{A10}$$

となるので，(A8) 式の分子は $n\sigma_{xy}$ となり，以上から $b = \dfrac{\sigma_{xy}}{\sigma_x^2}$ が得られます。

付録２：２次関数の最小化を用いた最小二乗法の導出

(1) 式を展開すると（以下，$\sum$ の添字を省略します),

$$\begin{aligned} L &= \sum_{i=1}^{n} \{y_i - (a + bx_i)\}^2 \\ &= b^2 \sum x_i^2 + 2ab \sum x_i + na^2 - 2b \sum x_i y_i - 2a \sum y_i + \sum y_i^2 \end{aligned} \tag{A11}$$

となります。ここで，$\sum x_i^2 = A, \sum x_i = B, \sum x_i y_i = C, \sum y_i = D, \sum y_i^2 = E$ とおくと，

$$L = Ab^2 \sum x_i^2 + 2Bab + na^2 - 2Cb - 2Da + E \tag{A12}$$

となり，a について整理すると

$$L = na^2 + 2(bB - D)a + (Ab^2 - 2Cb + E) \tag{A13}$$

となります。この式の第２項までを a の２次式とみて，平方完成を行うと

$$L = n\left(a + \frac{bB - D}{n}\right)^2 - \frac{(bB - D)^2}{n} + (Ab^2 - 2Cb + E) \tag{A14}$$

が得られます。この式の第２項以降を b について整理し，b の２次式とみて平方完成を行うと

$$\begin{aligned} &- \frac{(bB - D)^2}{n} + (Ab^2 - 2Cb + E) \\ &= (A - \frac{B^2}{n})b^2 - 2(C - \frac{BD}{n})b + (E - \frac{D^2}{n}) \\ &= \left(\frac{nA - B^2}{n}\right)\left(b - \frac{nC - BD}{nA - B^2}\right)^2 - \left(\frac{nA - B^2}{n}\right)\left(\frac{nC - BD}{nA - B^2}\right)^2 + (E - \frac{D^2}{n}) \end{aligned} \tag{A15}$$

となり，よって L は

$$\begin{aligned} L = n\left(a + \frac{bB - D}{n}\right)^2 + \left(\frac{nA - B^2}{n}\right)\left(b - \frac{nC - BD}{nA - B^2}\right)^2 \\ + \left[-\left(\frac{nA - B^2}{n}\right)\left(\frac{nC - BD}{nA - B^2}\right)^2 + (E - \frac{D^2}{n})\right] \end{aligned} \tag{A16}$$

のように，a, b 両方について平方の形になります。

ここで，$\dfrac{nA - B^2}{n}$ が正であれば，２つの平方の項はいずれも係数が正となります。よって，それらの項がいずれも 0，すなわち

$$\begin{aligned} a + \frac{bB - D}{n} = 0 \\ b - \frac{nC - BD}{nA - B^2} = 0 \end{aligned} \tag{A17}$$

のときに L が最小となり，この２つの式から a, b を求めることができます。そこで，$\dfrac{nA - B^2}{n}$ すなわち $A - \dfrac{B^2}{n}$ が正であることを確かめてみましょう。

$\sum x_i^2 = A, \sum x_i = B$ ですから，

$$\begin{aligned} A - \frac{B^2}{n} &= \sum x_i^2 - \frac{(\sum x_i)^2}{n} \\ &= n(\bar{x^2} - \bar{x})^2 \end{aligned} \tag{A18}$$

となります。付録 1 の (A9) 式で述べたように, x の分散を σ_x^2 とすると $\sigma_x^2 = \bar{x^2} - \bar{x}^2$ ですから, $\dfrac{nA - B^2}{n} = n\sigma_x^2 \geqq 0$ となります。

以上から，まず b について

$$b = \frac{nC - BD}{nA - B^2} \tag{A19}$$

が得られます。ここで，$\sum x_i^2 = A, \sum x_i = B, \sum x_i y_i = C, \sum y_i = D$ を代入すると，分母はさきほどの計算で $n^2\sigma_x^2$ となります。分子は

$$\begin{aligned} nC - BD &= n\sum x_i y_i - \sum x_i \sum y_i \\ &= n^2 \left(\frac{\sum x_i y_i}{n} - \frac{\sum x_i}{n} \cdot \frac{\sum y_i}{n} \right) \end{aligned} \tag{A20}$$

となります。ところで，x, y の共分散を σ_{xy} とすると，これもさきほどの (A10) 式で述べたように $\sigma_{xy} = \dfrac{\sum x_i y_i}{n} - \bar{x}\bar{y}$ なので，これを使うと分子は $n^2\sigma_{xy}$ となり，

$$b = \frac{n^2\sigma_{xy}}{n^2\sigma_x^2} = \frac{\sigma_{xy}}{\sigma_x^2} \tag{A21}$$

が得られます。

付録３：残差と決定係数の関係の導出

残差の定義から

$$\sum d_i^2 = \sum (y_i - \hat{y_i})^2 = \sum \{y_i - (bx_i + a)\}^2 \tag{A22}$$

で，さらに本文の (2) 式を使って，a を b で表すと

$$\begin{aligned} \sum d_i^2 &= \sum \{y_i - (bx_i + (\bar{y} - b\bar{x}))\}^2 \\ &= \sum [(y_i - \bar{y})^2 - 2b(y_i - \bar{y})(x_i - \bar{x}) + b^2(x_i - \bar{x})^2] \end{aligned} \tag{A23}$$

となります。ここで，付録 1 ・付録 2 で求めたように

$$b = \frac{\sigma_{xy}}{\sigma_x^2} \tag{A24}$$

ですから，これを代入すると

$$
\begin{aligned}
\sum d_i^2 &= \sum (y_i - \bar{y})^2 - 2\frac{\sigma_{xy}}{\sigma_x^2}\sum (y_i - \bar{y})(x_i - \bar{x}) + \left\{\frac{\sigma_{xy}}{\sigma_x^2}\right\}^2 \sum (x_i - \bar{x})^2 \\
&= \sum (y_i - \bar{y})^2 - 2\frac{\{\sum (x_i - \bar{x})(y_i - \bar{y})\}^2}{\sum (x_i - \bar{x})^2} + \frac{\{\sum (x_i - \bar{x})(y_i - \bar{y})\}^2}{\sum (x_i - \bar{x})^2} \\
&= \sum (y_i - \bar{y})^2 - \frac{\{\sum (x_i - \bar{x})(y_i - \bar{y})\}^2}{\sum (x_i - \bar{x})^2} \\
&= \sum (y_i - \bar{y})^2 - \frac{\{\sum (x_i - \bar{x})(y_i - \bar{y})\}^2}{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2} \sum (y_i - \bar{y})^2 \qquad \text{(A25)}
\end{aligned}
$$

となります。ここで相関係数の定義を用いると

$$
\sum d_i^2 = \sum (y_i - \bar{y})^2 - r_{xy}^2 \sum (y_i - \bar{y})^2 = (1 - r_{xy}^2) \sum (y_i - \bar{y})^2 \qquad \text{(A26)}
$$

が得られます。