

4 回帰直線

4.1 回帰分析の世界

こんな数式があります：

$$(\text{ゾウの表面積}) = -8.245 + 6.807 \times (\text{肩の高さ}) + 7.073 \times (\text{前足の太さ})$$

なんだこりゃ？ 『ゾウの表面積』って・・・一体何に使うんだ・・・

4.1.1 インド象の体表面積

基礎代謝量（BEE）とは、生命維持のために最低限必要なエネルギーのことで、24時間あたりの総エネルギー消費量のおよそ 60 %程度になります。主に医療の場面において、患者に必要な食事の量を決定したり、投与する薬の量を決定するときに大切な指標となります。

■基本的な計算式：（基礎代謝量）＝（基礎代謝基準値）×（体重）
基礎代謝基準値：18-29 歳 男性 24.0、女性 22.1

■ Harris-Benedict 方程式（日常の運動量が多い人には適さない）：

$$(\text{男性の基礎代謝}) = 66 + 13.7 \times (\text{体重}) + 5.0 \times (\text{身長}) - 6.8 \times (\text{年齢})$$

$$(\text{女性の基礎代謝}) = 665.1 + 9.6 \times (\text{体重}) + 1.7 \times (\text{身長}) - 7.0 \times (\text{年齢})$$

■国立健康・栄養研究所の計算式：

$$(\text{基礎代謝量}) = \{0.1238 + 0.0481 \times (\text{体重}) + 0.0234 \times (\text{身長}) - 0.0138 \times (\text{年齢}) - (\text{性別指数})\} \times \frac{1000}{4.186}$$

＊性別指数：0.5473（男性）、0.5473 × 2（女性）

■国立スポーツ科学センターの計算式：

$$(\text{基礎代謝量}) = 28.5 \times \{(\text{体重}) - (\text{脂肪量})\}$$

しかし、（ヒトも含めて）恒温動物の基礎代謝量は、体重よりもむしろその体表面積に近似的に比例することが知られています（M.Rubner 1883）。体表面積から基礎代謝量を計算し、動物への薬物投与量を決定したりしています。ちなみに日本人の基礎代謝量は、体表面積 1 平方メートル当り毎時 35～37 キロカロリーとされています。

ゾウの保護に取り組んでいるインドの研究者たちは、主に薬の投与量を決定するためにインド象の基礎代謝を知らなければなりません。基礎代謝が体表面積に比例するのは分かっている、体の表面積を計測するのはなかなか難しそうですね。ゾウの体の表面積を 1 頭 1 頭調べるのはものすごい手間ですし、計測中にゾウが暴れて死亡事故まであったとか。

そこで彼らは CG 技術を駆使して得た統計データを分析し、遂に比較的計測しやすい部位の数値から体表面積を『類推する』ための方程式を編み出しました（K.P.Sreekumar, G.Nirmalan 1990）。これがソレです：

■ Sreekumar 方程式：

$$(\text{体表面積}) = -8.245 + 6.807 \times (\text{肩までの高さ}) + 7.073 \times (\text{前足の周長})$$

ちなみに、この式はインド象限定で、アフリカ象には適さないとのこと。さらに言えば、彼らはこの研究によって 2002 年イグノーベル賞（数学賞）を受賞しました。

4.1.2 ボルドーワインの先物取引

仏ボルドー産の赤ワインは、主にカベルネ・ソーヴィニヨン種の葡萄で造られ、その特性上、造ってすぐに飲むのではなく、何年か（時には 10 年以上）熟成させてこそその真価が発揮されます。従ってワインは秋に収穫し、醸造されてから数年後ようやく販売されることになるのですが、ここに先物取引の余地が生じます。つまり、まだ熟成の途中の段階で将来の味を予測し、値付けし、買い取るという販売形態があるのです（プリムールと呼ばれています）。

どんな業者も、良いワイン（＝売れるワイン）はたくさん仕入れたいものです。しかし熟成が完成し評価が定まってからでは価格も高いし、ものすごい数の業者が群がるので思うような量は仕入れられません（高い金を払えば可能ですが・・・）。そこで、まだ評価が定まらない時期に早めに押さえてしまおうということですね。もちろんギャンブル的な要素が入り込んできます。つまり、（熟成後）良いワインになると思ったのに

思ったほどではなかった、とか。そうすると安く買い叩かれ、大損をすることにもなり兼ねません。もちろんそんなに良くなるとは思わず安い金額で仕入れたワインが熟成で大化けし、高価で買われていくこともあるわけですから難しい世界です。

そうなってくると人は、何とか現時点（醸造直後）の情報から10年後の価値を割り出したいと考えるでしょう。そんな数式があれば良いのにな、と。

あるんです。

■ Ashenfelter 方程式：

$$(\text{ワインの価値}) = (\text{前年の10月～3月の雨量}) \times 0.00117 - (\text{8,9月の雨量}) \times 0.00386 \\ + (\text{4～9月の平均気温}) \times 0.616 + (\text{熟成年数}) \times 0.02358 - 12.145$$

プリンストンの経済学部教授だった Orley Clark Ashenfelter が1984年に提唱した計算式です。ワイン好きだった彼は、ワインの価値が冬の雨量や育成期の気温が大きいほど高くなり、収穫期の雨量が少ないほど高くなることを統計データから導き出し、これらを合わせて重回帰分析と呼ばれる手法で計算式を完成させました。

ワイン評論家には最初酷評された方程式でしたが、驚くほど良く当たるということで一定の評価を得ることになりました。

4.2 既存データから未知の値を推測すること

建築用木材の原料となるある特定種の森林で、試験的に6本について木の高さと幹の太さを測定しました：

木の高さ (m)	8.7	6.8	7.9	7.0	7.1	6.1
幹の周長 (m)	0.75	0.55	0.72	0.61	0.66	0.58

得られたデータは木の高さと幹の周囲の長さがペアになった形（例えば (8.7, 0.75) など）をしており、このような形式のデータを2次元のデータと言うことにします。一般には更に高次元のデータを考える事もあります。

高さ8mの木が調達したい場合に、1本1本の木の高さをその都度測定するのは手間が掛かり過ぎるため、今回得られたサンプルデータからこの樹種の幹の太さ y と高さ x

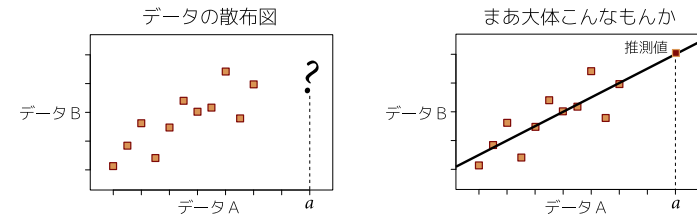
の間の（大雑把な）関係 $y = f(x)$ を見だし、 $x = 8$ を代入して高さが8mの木の幹の太さの概算値を求めておけば、幹の太さを測るだけで高さが8m（と思われる）の木を見つけることが出来るようになります。ここで

『木の高さが8mのときに木の幹の周長は何mか』

と云う問題が生じます。

一般に2次元のデータ (a_j, b_j) , $j = 1, 2, \dots, n$ が与えられ、第1成分データを A 、第2成分データを B として2次元のデータを (A, B) と書きます。この与えられたデータ (A, B) を基にして、『仮に A の値が a であった（原因）ならば B の値はどれくらい（結果）だろうか？』と云う問題を考えてみます。この問題において B のことを目的変数、 A を説明変数と言います。説明変数 A で目的変数 B を『説明する』と云う感じです。

横軸に A 、縦軸に B を取ってデータを2次元の点として可視化した図（これを散布図と言います）が下左図のようになっていましょう。こう云った場合に A と B の間の関係を探る最も単純な方法は、下右図の様に仮定の直線を図中に引いて『まあ、大体こんなもんだろ』と話を単純化してみる方法です（線形回帰法と言います）。



要するに一般にはデータ A とデータ B の間にはこの直線の示す1次関数の関係があるが、個体差のばらつきによって多少ぶれた値が出ているのだと空想すると云う事です。

4.3 回帰直線

問題はこの直線の引き方ですが、当然の要求として引いた直線の表す仮想値と実際のデータ（現実値）の差が出来るだけ小さくなる様にしたいわけです。具体的には与えられたデータ $\{(a_j, b_j)\}$ に対して、直線 $y = sx + t$ を考えて、 $B = sA + t$ と表せると仮定した場合の仮想値 $sa_j + t$ と現実値 b_j の差（＝ずれ）の自乗平均：

$$\frac{1}{n} \sum_{j=1}^n (sa_j + t - b_j)^2$$

が最小となる様な直線を求めます（これを最小自乗法と言います）。

差の自乗平均を具体的に計算すれば平方完成によって

$$\begin{aligned} & \frac{1}{n} \sum_{j=1}^n (a_j s + t - b_j)^2 \\ &= \frac{1}{n} \sum_{j=1}^n (a_j^2 s^2 + t^2 + b_j^2 + 2a_j s t - 2b_j t - 2a_j b_j s) \\ &= t^2 + 2E[A]st - 2E[B]t + E[A^2]s^2 + -2E[AB]s + E[B^2] \\ &= (t + E[A]s - E[B])^2 + \text{Var}[A] \left(s - \frac{E[AB] - E[A]E[B]}{\text{Var}[A]} \right)^2 + (\text{定数}) \end{aligned}$$

と書けますから（ただし $\text{Var}[A] \neq 0$ である必要があります）、

$$s = \frac{E[AB] - E[A]E[B]}{\text{Var}[A]}, \quad t = E[B] - E[A] \frac{E[AB] - E[A]E[B]}{\text{Var}[A]}$$

のときに最小値になります。ここに含まれている $E[AB] - E[A]E[B]$ のことを2次元のデータ (A, B) の共分散と言って記号 $\text{Cov}[A, B] = E[AB] - E[A]E[B]$ で表すことにすれば（意味は後で）、求める理想と現実の差が最小となる直線の方程式は

$$y = \frac{\text{Cov}[A, B]}{\text{Var}[A]}x + E[B] - \frac{E[A]\text{Cov}[A, B]}{\text{Var}[A]}$$

あるいは

$$y - E[B] = \frac{\text{Cov}[A, B]}{\text{Var}[A]}(x - E[A])$$

と書ける事が分かります。この直線を、 B の A への回帰直線（regression line）と言います。丁度『(平均値、平均値)』の点 $(E[A], E[B])$ を通っていますね。

問題 4.3.1 次の表は、ランダムに選ばれた6本の同一種の木について、高さ A と幹の周長 B を測定したものです（単位 m）。

A	8.7	6.8	7.9	7.0	7.1	6.1
B	0.75	0.55	0.72	0.61	0.66	0.58

(1) A, B の平均値、分散、共分散、 B の A への回帰直線を求めて下さい。

(2) 高さが 8.0 の木の幹の周囲を推測して下さい。

$$E[A] = \frac{43.6}{6}, \quad E[B] = \frac{3.87}{6}$$

$$\text{Var}[A] = E[A^2] - \{E[A]\}^2 = \frac{320.96}{6} - \left(\frac{43.6}{6}\right)^2 = \frac{24.8}{36}$$

$$\text{Var}[B] = \frac{2.5275}{6} - \frac{14.9769}{36} = \frac{0.1881}{36}$$

$$\text{Cov}[A, B] = E[AB] - E[A]E[B] = \frac{28.447}{6} - \frac{168.732}{36} = \frac{1.95}{36}$$

従って B の A への回帰直線は

$$y - \frac{3.87}{6} = \frac{1.95}{24.8} \left(x - \frac{43.6}{6} \right)$$

となり、 $x = 8$ を代入すれば高さが 8m の時の幹の周長の推測値が得られます：

$$y = \frac{1.95 \cdot 8 \cdot 6 - 1.95 \cdot 43.6 + 3.87 \cdot 24.8}{24.8 \cdot 6} = \frac{104.556}{148.8} \approx 0.70266 \approx 0.70.$$

従って幹の周長が 0.70m の木を切れば長さ 8m 程度の木が得られると推察されます。□

4.4 共分散

2次元データ (A, B) に対して、2つの成分を加えて得られる1次元データを考えます。

元データ $(A, B) : (a_1, b_1), (a_2, b_2), \dots, (a_n, b_n)$

↓ 加工

派生データ $A + B : a_1 + b_1, a_2 + b_2, \dots, a_n + b_n$

すると和 $A + B$ の平均値と分散は、

$$E[A + B] = \frac{1}{n} \sum_{j=1}^n (a_j + b_j) = \frac{1}{n} \sum_{j=1}^n a_j + \frac{1}{n} \sum_{j=1}^n b_j = E[A] + E[B]$$

$$\text{Var}[A + B] = E[(A + B - E[A + B])^2]$$

$$= \frac{1}{n} \sum_{j=1}^n (a_j + b_j - E[A] - E[B])^2$$

$$= \frac{1}{n} \sum_{j=1}^n \{(a_j - E[A]) + (b_j - E[B])\}^2$$

$$= \frac{1}{n} \sum_{j=1}^n (a_j - E[A])^2 + 2 \frac{1}{n} \sum_{j=1}^n (a_j - E[A])(b_j - E[B]) + \frac{1}{n} \sum_{j=1}^n (b_j - E[B])^2$$

$$= \text{Var}[A] + 2E[(A - E[A])(B - E[B])] + \text{Var}[B]$$

と計算され、和の平均は平均の和ですが、和の分散は単純に分散の和とはならずクロスタームが入って来る事になります。このクロスタームは、

$$\begin{aligned} E[(A - E[A])(B - E[B])] \\ &= \frac{1}{n} \sum_{j=1}^n (a_j b_j - E[A] b_j - E[B] a_j + E[A] E[B]) \\ &= \frac{1}{n} \sum a_j b_j - E[A] \frac{1}{n} \sum b_j - \left(\frac{1}{n} \sum a_j \right) E[B] + E[A] E[B] \\ &= E[AB] - E[A] E[B] \end{aligned}$$

と変形され、これこそまさにさっきの共分散である事が分かります。それぞれの平均からのずれの積の平均と云うこの形を定義とするのが普通です

定義 4.4.1 2次元のデータ (A, B) に対して次の右辺が存在するとき：

$$\text{Cov}[A, B] = E[(A - E[A])(B - E[B])]$$

これを A, B の共分散 (covariance) と言います。

有限データだけでなく無限データの場合にも同様に定義されます。また、この定義の形だけでなく、さっきの $E[AB] - E[A]E[B]$ の形の方が計算しやすい場合があるので覚えておくべきです。ちなみに対称性： $\text{Cov}[A, B] = \text{Cov}[B, A]$ は自明ですね。

4.5 相関係数

定義 4.5.1 2次元のデータ (A, B) に対して、 $\text{Var}[A], \text{Var}[B]$ が存在して共に0でないとき共分散を標準偏差の積で割ったもの：

$$\text{Cor}[A, B] = \frac{\text{Cov}[A, B]}{\sqrt{\text{Var}[A]} \sqrt{\text{Var}[B]}}$$

をこれらのデータの相関係数 (correlation coefficient) と言います。

有限データの場合に相関係数の定義を詳しく見ると、

$$\begin{aligned} (A, B \text{ の相関係数}) &= \frac{\frac{1}{n} \sum_{j=1}^n (a_j - E[A])(b_j - E[B])}{\sqrt{\frac{1}{n} \sum_{j=1}^n (a_j - E[A])^2} \sqrt{\frac{1}{n} \sum_{j=1}^n (b_j - E[B])^2}} \\ &= \frac{\sum_{j=1}^n (a_j - E[A])(b_j - E[B])}{\sqrt{\sum_{j=1}^n (a_j - E[A])^2} \sqrt{\sum_{j=1}^n (b_j - E[B])^2}} \end{aligned}$$

であって、これは2つの n 次元ベクター

$$\mathbf{v} = (a_1 - E[A], \dots, a_n - E[A]), \mathbf{w} = (b_1 - E[B], \dots, b_n - E[B])$$

の内積をそれらのベクターの大きさの積で割ったものになっています。

従って内積が、それら2本のベクターの成す角を θ とした時に $\mathbf{v} \cdot \mathbf{w} = |\mathbf{v}| |\mathbf{w}| \cos \theta$ と表される事から考えれば、相関係数とは、正にその2つのデータの平均からのずれを新たなデータとして考え、それらの“成す角”の cosine であることが分かります。

内積を見ていたのではベクターの大きさが影響していて角度は分かりにくくなっていますので、ベクターの大きさを割って正規化してやる事によって2つのベクターがどんな幾何学的関係にあるかが分かり易くなるわけです。

従って、相関係数は -1 から 1 までの値をとり、それぞれのデータの平均値との差をとったものがベクターとして見て平行で同じ向きならば相関係数は 1 、平行で逆向きならば相関係数は -1 となる事が分かります。

2つのデータが平行であると言う事は、要するに一方のデータが他方のデータの定数倍になっていると云うことですが、定数を付加しても分散は変わらない事を考慮すれば、2つのデータ A, B が線形関係 $B = pA + w$ を満たす時に、 $p > 0$ なら相関係数は 1 、 $p < 0$ なら相関係数は -1 となるのだと云うことが分かります。また、当然、直交していれば相関係数は 0 です。

共分散はそれぞれのデータ値の大きさなどを反映しますので2つのデータの『線形関係そのもの』を見るのには適していません。そこで共分散を正規化して2つのデータの関係にフォーカスしたものが相関係数です。

しかしあくまでも相関係数は2つのデータの間の線形関係を見ているだけですので、非線形な関係がある場合には『2つのデータの間に強い関連性があるにも拘らず相関係数は比較的小さくなる』事もあります。この点には十分注意して相関係数と云うものを理解する必要があります。

4.6 問題演習

基本演習 4.1 $Cov[A, A] = Var[A]$ となる事を確かめて下さい。

基本演習 4.2 任意の定数 v, w に対して $Cov[A, vB + w] = vCov[A, B]$ となる事を確かめて下さい。

基本演習 4.3 先の問題 4.3.1 において A の B への回帰直線を求めて下さい。

ただし、あくまで A を x -軸、 B を y -軸に対応させて $x = vy + w$ の形の直線の方程式を求めて下さい。

基本演習 4.4 ある日のプロ野球公式戦において 12 球団の安打数と得点は以下の表の通りでした。得点データを A 、安打数データを B として以下の問いに答えて下さい。

得点 A	3	2	7	1	1	2	1	6	3	0	7	3
安打 B	5	8	6	5	8	8	7	11	7	4	7	8

(1) A, B それぞれの平均値 $E[A], E[B]$ と A の分散 $Var[A]$ を求めて下さい。

(2) B の A への回帰直線を求め、得点が 5 点である時の安打数を概算して下さい。

基本演習 4.5 次の表はある地域で産出される赤ワインについて、年度ごとの育成期の平均気温 D 、10 年後の市場価値（1961 年産の価格との比の対数） P をまとめたものです：

	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969
D	16.5	17.3	16.3	15.7	17.3	15.4	16.5	16.2	16.2	16.6
P	-2.0	0	-1.1	-1.8	-1.2	-2.2	-0.7	-1.7	-2.2	-2.1

(1) D, P の平均値 $E[D], E[P]$ を計算して下さい。値は一切の近似をせずに求めて下さい。

(2) D の分散 $Var[D]$ と D と P の共分散 $Cov[D, P]$ を求めて下さい。値は一切の近似をせずに求めて下さい。

(3) P の D への回帰直線が

$$y - E[P] = \frac{Cov[D, P]}{Var[D]}(x - E[D])$$

となることを使って平均気温 D が 16.0 だった場合の価値 P の推定値を求めて下さい。

基本演習 4.6 ある洋菓子チェーン店はすべての店舗が駅前に立地しています。全 15 店舗の売上高 (万円／月) と、それぞれの駅の平均乗降客数 (十人／日) を調査したところ、次の結果を得ました。

乗降客数 A	148	186	279	179	216	127	152	196
売上高 B	129	194	217	165	162	154	203	269
乗降客数 A	126	78	211	259	255	115	173	
売上高 B	205	127	260	249	251	158	167	

売上高 B の、乗降客数 A への回帰直線を求め、乗降客数が 200 人の場合の売上高の予想値を求めてください。

発展演習 4.7 2次元のデータ (A, B) が与えられ、各成分データは平均値と分散をもち、分散はいずれも正であるとします。

このとき A, B それぞれを標準化したものを A^*, B^* と置く時、 $A^* \pm B^*$ の分散を計算する事により $|Cor[A, B]| \leq 1$ である事を証明して下さい。

発展演習 4.8 赤いサイコロと青いサイコロが 1 つずつあります。これらを同時に振って出た目をそれぞれ $R(\text{Red})$ 、 $B(\text{Blue})$ で表します。

このとき R と B の共分散を求めて下さい。

発展演習 4.9 次の 4 つの 2 次元データで平均値、分散、共分散を計算してください。

A_1	10.0	8.0	13.0	9.0	11.0	14.0	6.0	4.0	12.0	7.0	5.0
B_1	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68

A_2	10.0	8.0	13.0	9.0	11.0	14.0	6.0	4.0	12.0	7.0	5.0
B_2	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74

A_3	10.0	8.0	13.0	9.0	11.0	14.0	6.0	4.0	12.0	7.0	5.0
B_3	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73

A_4	8.0	8.0	8.0	8.0	8.0	8.0	8.0	19.0	8.0	8.0	8.0
B_4	6.58	5.76	7.71	8.84	8.47	7.04	5.25	12.50	5.56	7.91	6.89

発展演習 4.10 B の A への回帰直線が $y = sx + t$ であるとき、

$$\frac{E[(sA + t - E[B])^2]}{E[(B - E[B])^2]} = Cor[A, B]^2$$

であることを示してください。

問題演習解答

基本演習 4.1

$$\text{Cov}[A, A] = E[(A - E[A])(A - E[A])] = \text{Var}[A].$$

基本演習 4.2

$$\begin{aligned}\text{Cov}[A, vB + w] &= E[(A - E[A])(vB + w - E[vB + w])] \\ &= E[(A - E[A])(vB - vE[B])] \\ &= v\text{Cov}[A, B].\end{aligned}$$

基本演習 4.3 $x = 10.37y + 0.580$ (計算略) .基本演習 4.4 (1) $E[A] = 3, E[B] = 7, \text{Var}[A] = \frac{16}{3} \approx 5.33$

(2) $\text{Cov}[A, B] = \frac{17}{12}$ なので回帰直線は $y - 7 = \frac{17}{64}(x - 3)$ 。安打数 7.53 本。

基本演習 4.5

- (1) $E[D] = 16.4 \quad E[P] = -1.5$
 (2) $\text{Var}[D] = 0.326 \quad \text{Cov}[D, P] = 0.258$
 (3) 回帰直線は

$$y + 1.5 = \frac{0.258}{0.326}(x - 16.4)$$

ですから、 $x = 16.0$ を代入すると $y \approx -1.8166$ 。

基本演習 4.6

$$\begin{aligned}E[A] &= \frac{148 + 186 + \cdots + 173}{15} = \frac{2700}{15} = 180 \\ E[B] &= \frac{129 + 194 + \cdots + 167}{15} = \frac{2910}{15} = 194 \\ \text{Var}[A] &= \frac{(148 - 180)^2 + (186 - 180)^2 + \cdots + (173 - 180)^2}{15} = \frac{46428}{15} = 3095.2\end{aligned}$$

$A - 180$	-32	6	99	-1	36	-53	-28	16
$B - 194$	-65	0	23	-29	-32	-40	9	75
$A - 180$	-54	-102	31	79	75	-65	-7	
$B - 194$	11	-67	66	55	57	-36	-27	

$$\begin{aligned}\text{Cov}[A, B] &= \frac{(148 - 180)(129 - 194) + \cdots + (173 - 180)(167 - 194)}{15} \\ &= \frac{25737}{15} = 1715.8\end{aligned}$$

以上から回帰直線は以下の通り：

$$y - 194 = \frac{1715.8}{3095.2}(x - 180)$$

ここで $x = 200$ とすれば

$$y = \frac{1715.8}{3095.2} \cdot 20 + 194 \approx 205.1$$

です (きちんと計算していないのでミスがあるかも)。

発展演習 4.7 略。

発展演習 4.8 0.

発展演習 4.9

平均値、分散、積の平均値を計算しておきます：

$$E[A_1] = E[A_2] = E[A_3] = E[A_4] = 9$$

$$E[B_1] \approx E[B_2] \approx E[B_3] \approx E[B_4] \approx 7.5$$

$$\text{Var}[A_1] = \text{Var}[A_2] = \text{Var}[A_3] = \text{Var}[A_4] = 10$$

$$\text{Var}[B_1] \approx \text{Var}[B_2] \approx \text{Var}[B_3] \approx \text{Var}[B_4] \approx 3.75$$

$$\begin{aligned}E[A_1 B_1] &\approx \frac{10.0 \cdot 8.04 + 8.0 \cdot 6.95 + \cdots + 5.0 \cdot 5.68}{11} = \frac{797.6}{11} \approx 72.5 \\ E[A_2 B_2] &\approx \frac{10.0 \cdot 9.14 + 8.0 \cdot 8.14 + \cdots + 5.0 \cdot 4.74}{11} = \frac{797.59}{11} \approx 72.5 \\ E[A_3 B_3] &\approx \frac{10.0 \cdot 7.46 + 8.0 \cdot 6.77 + \cdots + 5.0 \cdot 5.73}{11} = \frac{797.47}{11} \approx 72.5 \\ E[A_4 B_4] &\approx \frac{8.0 \cdot 6.58 + 8.0 \cdot 5.76 + \cdots + 8.0 \cdot 6.89}{11} = \frac{797.58}{11} \approx 72.5\end{aligned}$$

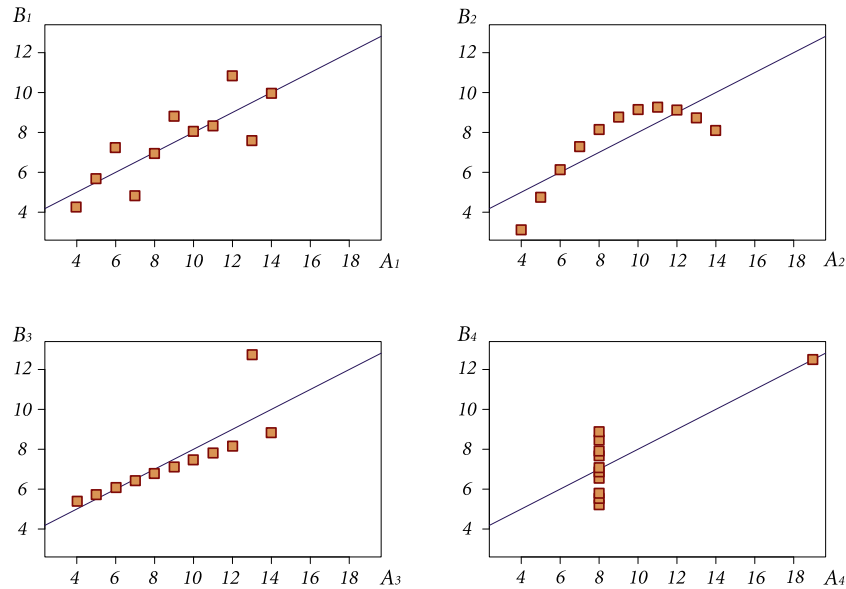
A_j, B_j それぞれのグループ内で平均値、分散が等しく、更に積の平均値も全て等しいので、共分散は全て等しくなります：

$$\text{Cov}[A_j, B_j] \approx 72.5 - 9 \cdot 7.5 \approx 5.0$$

従って4つのケースとも同一の回帰直線が得られます：

$$y \approx 0.5x + 3.0.$$

データそのものを散布図で見る限りはこれら4つの2次元データは相当に異なるタイプの分布をしていますが（下図）、平均値、分散、共分散などは全く同じであり、回帰直線も同一です。データそのものを表現すると云う目的に於いては、これらの基本的な統計量が如何に不十分なものであるか、また、線形回帰分析と云うものが如何に一面的なものか認識する良い機会でしょう。



発展演習 4.10

現実と平均値のズレの自乗平均のうち、理想推定値で説明できる部分の比は、

$$\frac{E[(sA + t - E[B])^2]}{E[(B - E[B])^2]}$$

ですが、 s, t の決め方によれば $E[B] = sE[A] + t$ でしたから

$$\begin{aligned} &= \frac{E[(sA + t - sE[A] - t)^2]}{E[(B - E[B])^2]} \\ &= \frac{s^2 \text{Var}[A]}{\text{Var}[B]} \end{aligned}$$

です。またここで $s = \frac{\text{Cov}[A, B]}{\text{Var}[A]}$ だったので

$$\begin{aligned} &= \frac{\text{Cov}[A, B]^2}{\text{Var}[A]\text{Var}[B]} \\ &= \text{Cor}[A, B]^2 \end{aligned}$$

と書けることが分かります。

□