

4 回帰直線

4.1 回帰分析の世界

こんな数式があります：

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

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、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。

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でした。そこで彼らは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：

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

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。

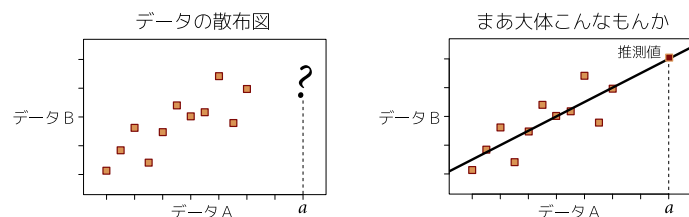
■ Ashenfelter 方程式：

$$(\text{ワインの価値}) = (\text{前年の10月～3月の雨量}) \times 0.00117 - (8,9月の雨量) \times 0.00386 + (4\sim9月の平均気温) \times 0.616 + (\text{熟成年数}) \times 0.02358 - 12.145$$

4.2 既存データから未知の値を推測すること

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

データから幹の太さと高さの間の関係を見だし、太さを測れば高さが分かるようにしたい。そのために『仮に A の値（高さ）が 8m であったならば B の値（幹の周長）はどれくらいだろうか？』を考える。



4.3 回帰直線

$B = sA + t$ と表せると仮定した場合の仮想値 $sa_j + t$ と現実値 b_j の差の自乗平均：

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

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）。

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

のときに最小値。共分散： $\text{Cov}[A, B] = E[AB] - E[A]E[B]$ を使えば、

$$y - E[B] = \frac{\text{Cov}[A, B]}{\text{Var}[A]}(x - E[A]) \quad (B \text{ の } A \text{ への回帰直線})$$

問題 4.3.1 (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。

$$E[A] = \frac{43.6}{6}, \quad E[B] = \frac{3.87}{6}, \quad \text{Var}[A] = \frac{24.8}{36}, \quad \text{Var}[B] = \frac{0.1881}{36}$$

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

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、これに $x = 8$ を代入すれば

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

ですから、幹の周長が 0.70 メートルの木を切れば長さ 8 メートル程度の木が得られると推察されます。□

4.4 共分散

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

$$E[(A - E[A])(B - E[B])] = E[AB] - E[A]E[B]$$

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。

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。

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