

『AI・データサイエンスのための数学入門』学習補助資料

相関・回帰・決定係数をつなぐ

散布図から、計算とモデルの評価へ

1 相関係数は「偏差ベクトルの向き」を測る

計算を見通しよくするため、次の記号を使います。

$$S_{xx} = \sum (x_i - \bar{x})^2, \quad S_{yy} = \sum (y_i - \bar{y})^2, \quad S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}).$$

相関係数は $r = S_{xy} / \sqrt{S_{xx}S_{yy}}$ です。 $S_{xx}, S_{yy} > 0$ を仮定します。一方の変数が一定なら、分母が 0 となり、相関係数は定義できません。

なぜ $-1 \leq r \leq 1$ なのか

偏差ベクトルを $\mathbf{u} = (x_i - \bar{x})_{i=1}^n$ 、 $\mathbf{v} = (y_i - \bar{y})_{i=1}^n$ とすると、

$$r = \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}.$$

コーシー＝シュワルツの不等式 $|\mathbf{u}^\top \mathbf{v}| \leq \|\mathbf{u}\| \|\mathbf{v}\|$ から範囲が分かります。幾何学的には、2 つの偏差ベクトルのなす角の余弦です。

$r = 1$ または -1 になるのは、 $y_i - \bar{y} = c(x_i - \bar{x})$ となる非零定数 c があるときです。すなわち、全点が傾きの正または負の 1 本の直線上にあります。

相関と傾きは違う

同じデータについて、切片付き最小二乗回帰の傾きは

$$a = \frac{S_{xy}}{S_{xx}} = r \sqrt{\frac{S_{yy}}{S_{xx}}} = r \frac{s_y}{s_x}$$

です。 s_x, s_y の分母は同じものを用います。相関は単位を持ちませんが、傾きは「 y の単位 / x の単位」を持ちます。

単位を変えると何が変わるか

$x' = 100x$ と単位を変えると、相関係数は変わらず、回帰の傾きは $a/100$ になります。一般に、正の倍率と定数の加算は相関係数を変えません。一方だけを負の数倍すると符号が逆転します。

2 最小二乗法の公式を、平方完成で導く

説明変数が一定でない、すなわち $S_{xx} > 0$ とします。 $u_i = x_i - \bar{x}$ 、 $v_i = y_i - \bar{y}$ 、 $d = b + a\bar{x} - \bar{y}$ とおくと、

$$y_i - (ax_i + b) = v_i - au_i - d.$$

$\sum u_i = \sum v_i = 0$ なので、展開した交差項の一部が消え、

$$\begin{aligned} Q(a, b) &= \sum (v_i - au_i - d)^2 \\ &= S_{yy} - 2aS_{xy} + a^2S_{xx} + nd^2 \\ &= S_{xx} \left(a - \frac{S_{xy}}{S_{xx}} \right)^2 + n(b + a\bar{x} - \bar{y})^2 + S_{yy} - \frac{S_{xy}^2}{S_{xx}}. \end{aligned}$$

最初の 2 項は非負です。両方が 0 となるところで最小になるため、

$$a = \frac{S_{xy}}{S_{xx}}, \quad b = \bar{y} - a\bar{x}$$

を得ます。これは微分を使わずに、最小性と一意性まで示す証明です。

最小化しているのは「縦方向の差」

通常の回帰では、各 x_i を固定し、 y_i と $ax_i + b$ の差を測ります。点から直線への最短距離を最小化しているわけではありません。

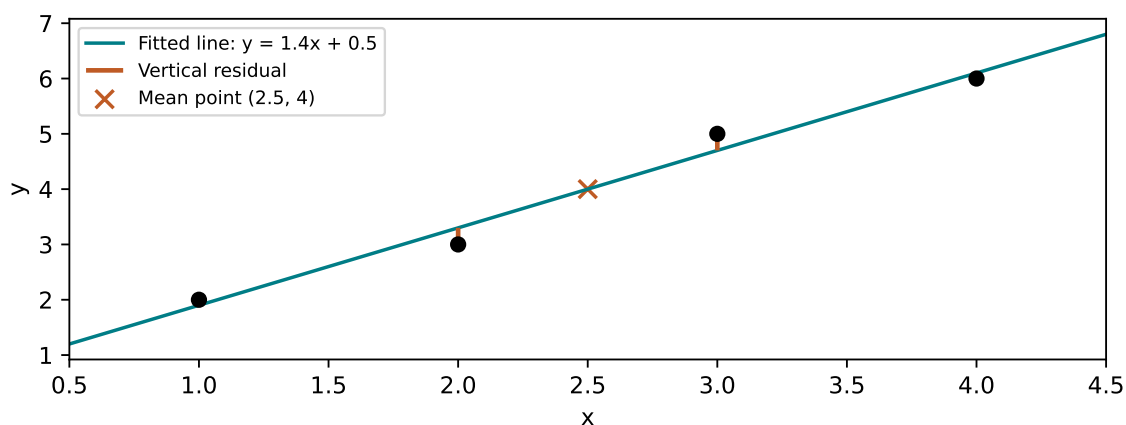


図 1：本文の練習問題のデータ。橙の縦線が残差を表します。回帰直線は平均点 (2.5, 4) を通ります。

逆向きの回帰は、逆関数ではない

y から x を予測すると、傾きは S_{xy}/S_{yy} です。 x から y を予測する傾きとの積は r^2 なので、 $|r| < 1$ なら互いの逆数にはなりません。

3 残差の直交性から $R^2 = r^2$ へ

切片を含む最小二乗回帰の残差を $e_i = y_i - \hat{y}_i$ とすると、

$$\sum e_i = 0, \quad \sum (x_i - \bar{x})e_i = 0$$

が成り立ちます。実際、 $e_i = (y_i - \bar{y}) - a(x_i - \bar{x})$ に代入すれば、後者は $S_{xy} - aS_{xx} = 0$ です。

全変動を 2 つに分ける

$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + e_i$ を平方して足します。 $\hat{y}_i - \bar{y} = a(x_i - \bar{x})$ より交差項が消え、

$$\underbrace{\sum (y_i - \bar{y})^2}_{\text{SST:}} = \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{\text{SSR:}} + \underbrace{\sum e_i^2}_{\text{SSE:}}.$$

この等式は、偏差ベクトルを直交する 2 成分に分けたピタゴラスの定理と見ることができます。

$S_{yy} > 0$ なら、

$$R^2 = 1 - \frac{\text{SSE}}{S_{yy}} = \frac{a^2 S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx} S_{yy}} = \boxed{r^2}.$$

「80% を説明」の基準は何か

全員に平均値 $\bar{y}$ を予測するモデルの二乗誤差は SST です。 $R^2 = 0.8$ は、**同じデータ上で回帰モデルがこの基準の二乗誤差を 80% 減らしたことを意味します。**「80% の人に正しく予測できた」ではありません。

成立する条件も覚える

$0 \leq R^2 \leq 1$ と r^2 との一致は、ここでは**切片付き単回帰を最小二乗法で当てはめた学習データ上の性質**です。切片なしの場合や、別の評価データでは、そのまま成り立ちません。 y が一定なら、この定義の R^2 は分母 0 で定義できません。

評価データで R^2 が負になる例

新しい観測値が (0, 2)、モデルの予測が (3, 3) なら、評価データの平均は 1 です。SST = 2、SSE = 10 なので $R^2 = -4$ です。この評価データでは、平均 1 を予測する基準より誤差が大きかったと分かります。

4 散布図と残差で、数値だけでは見えない関係を読む

無相関でも、完全な関係がありうる

$x = (-2, -1, 0, 1, 2)$ 、 $y = x^2 = (4, 1, 0, 1, 4)$ とします。 $\bar{x} = 0$ 、 $\bar{y} = 2$ で、

$$S_{xy} = (-2) \cdot 2 + (-1) \cdot (-1) + 0 \cdot (-2) + 1 \cdot (-1) + 2 \cdot 2 = 0.$$

したがって $r = 0$ です。しかし y は x から完全に決まります。これら 5 点から等確率で 1 点を選ぶ確率モデルでも、2 変数は無相関ですが独立ではありません。

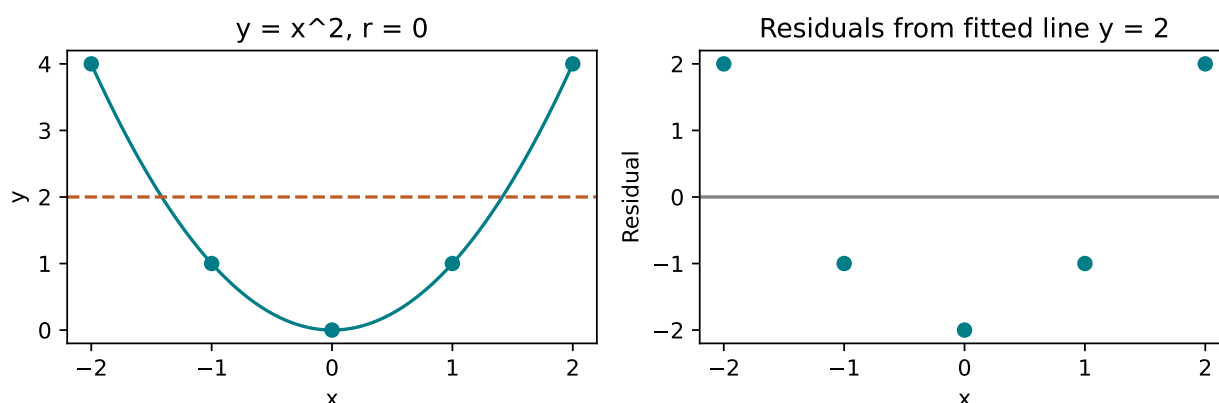


図 2：このデータの単回帰直線は $\hat{y} = 2$ 。残差は $(2, -1, -2, -1, 2)$ となり、U 字型の規則が残ります。二次項の導入を考える手がかりです。

残差図で見る代表的な形

見える形	確認したいこと
曲線・U 字型	直線モデルでは関係を捉えきれていない可能性。
右へ広がる扇形	誤差のばらつきが x により変わる可能性。
時刻順の波・連続した偏り	誤差の時間的な依存や、未考慮の季節性。
少数の大きな残差	測定ミス、特殊な条件、モデル不足など。

外れ値・外挿・因果関係

x 方向に他から離れた点は、残差が小さくても回帰直線を強く左右することがあります。除外を先に決めず、値の妥当性と、その点を含む場合・含まない場合を比較します。

観測した x の範囲外への予測は**外挿**です。良い当てはまりでも、範囲外で直線関係が続く保証はありません。また、気温が飲料売上と水辺の来訪者数の両方を増やすように、第三の変数による相関もあります。回帰の向きを決めても、因果関係の証明にはなりません。

5 本文 3 問を、1 つの計算表で解く

データ $(1, 2), (2, 3), (3, 5), (4, 6)$ について、 $\bar{x} = 2.5$ 、 $\bar{y} = 4$ です。

x_i	y_i	$u_i = x_i - 2.5$	$v_i = y_i - 4$	$u_i v_i$	u_i^2	v_i^2
1	2	-1.5	-2	3	2.25	4
2	3	-0.5	-1	0.5	0.25	1
3	5	0.5	1	0.5	0.25	1
4	6	1.5	2	3	2.25	4
		合計		7	5	10

問 1：相関係数

$S_{xy} = 7$ 、 $S_{xx} = 5$ 、 $S_{yy} = 10$ より、

$$r = \frac{7}{\sqrt{50}} \approx 0.98995.$$

この 4 点には強い正の線形関係があります。ただし、少数の観測から母集団全体の関係や因果関係まで確定することはできません。

問 2：回帰直線

$$a = \frac{7}{5} = 1.4, \quad b = 4 - 1.4 \cdot 2.5 = 0.5, \quad \boxed{\hat{y} = 1.4x + 0.5}.$$

x が 1 単位増えると、予測値が 1.4 単位増えます。個々の実測値が必ず 1.4 増えるという意味ではありません。

問 3：残差と決定係数

x_i	1	2	3	4
$\hat{y}_i$	1.9	3.3	4.7	6.1
$e_i = y_i - \hat{y}_i$	0.1	-0.3	0.3	-0.1
e_i^2	0.01	0.09	0.09	0.01

$$\text{SSE} = 0.20, \quad \boxed{R^2 = 1 - \frac{0.20}{10} = 0.98 = r^2}.$$

このデータ上で、平均 4 を一律に予測する場合の二乗誤差を 98% 減らしています。新しいデータでも同じ精度になるとは限りません。

6 重回帰：複数の方向への射影【補足】

重回帰の予測ベクトルは $\hat{Y} = X\hat{\beta}$ です。これは X の各列の線形結合なので、列ベクトルが張る空間に属します。最小二乗法は、その空間の中で Y に最も近い点を選びます。

正規方程式の導出

任意の方向 h に係数を動かすと、残差 $e = Y - X\hat{\beta}$ に対して、

$$Q(\hat{\beta} + th) = \|e\|^2 - 2th^\top X^\top e + t^2 \|Xh\|^2.$$

$t = 0$ で最小になるには一次項が全方向で 0 となる必要があります、

$$X^\top e = 0 \iff X^\top X\hat{\beta} = X^\top Y$$

を得ます。残差はすべての説明変数の列に直交します。 X の列が線形独立なら、 $X^\top X$ は可逆で係数が一意に定まります。

共線性で、何が決まらなくなるか

すべての観測で $x_2 = 2x_1$ なら、

$$\beta_1 x_1 + \beta_2 x_2 = (\beta_1 + 2\beta_2)x_1.$$

例えば $(\beta_1, \beta_2) = (3, 0)$ と $(1, 1)$ は同じ予測を与えます。係数を個別に識別できません。完全な共線性がなくても、それに近ければ係数がデータの小さな変化に敏感になります。

係数の読み方と、因果の区別

β_j は、他の説明変数を固定したモデル上で、 x_j を 1 単位増やしたときの平均的な予測値の差です。現実にはその変数だけを操作した因果効果とは、追加の仮定なしには一致しません。

当てはめと、統計的推測は別の段階

最小二乗解の計算自体に誤差の正規性は不要です。一方、係数の不偏性には典型的に $E[\epsilon | X] = 0$ 、通常の標準誤差には等分散・無相関など、正確な有限標本の t 検定にはさらに正規性などを仮定します。

同じ学習データで説明変数を追加すると、候補の予測が増えるため、残差平方和は増えず、通常の R^2 は下がりにません。変数選択には、評価用データや交差検証なども用いて予測性能を確かめます。

7 追加練習問題

問 1：単位の変更 データの回帰直線が $\hat{y} = 1.4x + 0.5$ 、相関係数が $r = 0.99$ だった。 $x' = 100x$ 、 $y' = y + 10$ と変換したとき、回帰直線と相関係数を求めよ。

問 2：逆向きの予測 本文のデータについて、 x を y から予測する回帰直線 $\hat{x} = cy + d$ を求めよ。 $\hat{y} = 1.4x + 0.5$ を x について解いた式との違いを説明せよ。

問 3：完全な負の相関 $(x, y) = (1, 7), (2, 5), (3, 3)$ について、 r 、回帰直線、 R^2 を求めよ。 R^2 だけで増減の向きが分かるか。

問 4：残差の合計 切片付き最小二乗回帰で、残差が $(1, -2, 1)$ だった。残差の合計と残差平方和を求め、「残差の合計が 0 なら完全に当てはまる」が誤りであることを説明せよ。

問 5：デザイン行列 3 件の観測について説明変数が $(x_1, x_2) = (1, 2), (2, 4), (3, 6)$ だった。切片を含むデザイン行列を書き、列の独立性と係数の一意性を判定せよ。

問 6：モデル評価 同じ学習データで、モデル A の R^2 は 0.80、説明変数を追加したモデル B の R^2 は 0.85 だった。「B は新しいデータにも必ず優れる」といえるか。理由を述べよ。

解答を読む前の確認

- 相関は対称、回帰は予測する向きがある。
- 回帰直線は平均点を通り、残差の合計は 0 になる。
- R^2 は大きさの指標で、傾きの符号は表さない。
- 学習データへの当てはまりと、未知データへの予測性能は区別する。

8 追加問題の解答・解説

問 1：単位の変更

$x = x'/100$ より、 $\hat{y}' = 1.4(x'/100) + 0.5 + 10$ です。したがって

$$\hat{y}' = 0.014x' + 10.5, \quad r' = 0.99.$$

正の単位変換や原点の移動では、偏差ベクトルのなす角は変わりません。

問 2：逆向きの予測

$c = S_{xy}/S_{yy} = 7/10 = 0.7$ 、 $d = 2.5 - 0.7 \cdot 4 = -0.3$ なので

$$\hat{x} = 0.7y - 0.3.$$

元の直線を逆に解くと $x = (y - 0.5)/1.4$ で、傾きは約 0.7143 です。最小化する誤差の方向が異なるため、同じ直線にはなりません。

問 3：完全な負の相関

全点が $y = -2x + 9$ にあるため、 $r = -1$ 、 $\hat{y} = -2x + 9$ 、 $R^2 = 1$ です。 R^2 では相関の符号が失われるため、正負は判断できません。

問 4：残差の合計

合計は $1 - 2 + 1 = 0$ ですが、残差平方和は $1 + 4 + 1 = 6$ です。正負が打ち消し合っても、各点でのずれは残ります。完全な当てはまりには、すべての残差が 0 であることが必要です。

問 5：デザイン行列

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 3 & 6 \end{pmatrix}.$$

第 3 列は第 2 列の 2 倍なので列は線形従属、階数は 2 です。 $\mathbf{X}^\top \mathbf{X}$ は可逆でなく、係数は一意ではありません。ただし、最小二乗の当てはめ値は列空間への直交射影として一意です。

問 6：モデル評価

必ず優れるとはいえません。変数追加により学習データへの当てはまりは良くなっても、偶然の変動まで取り込む過学習が起こりえます。同じ条件で評価用データの誤差を比べるなど、未知データへの性能を確認します。

講義の到達目標

散布図を読み、 S_{xx}, S_{xy}, S_{yy} から r と回帰直線を計算し、残差と R^2 を説明できること。さらに、無相関・非因果・外挿・過学習の違いを意識できること。