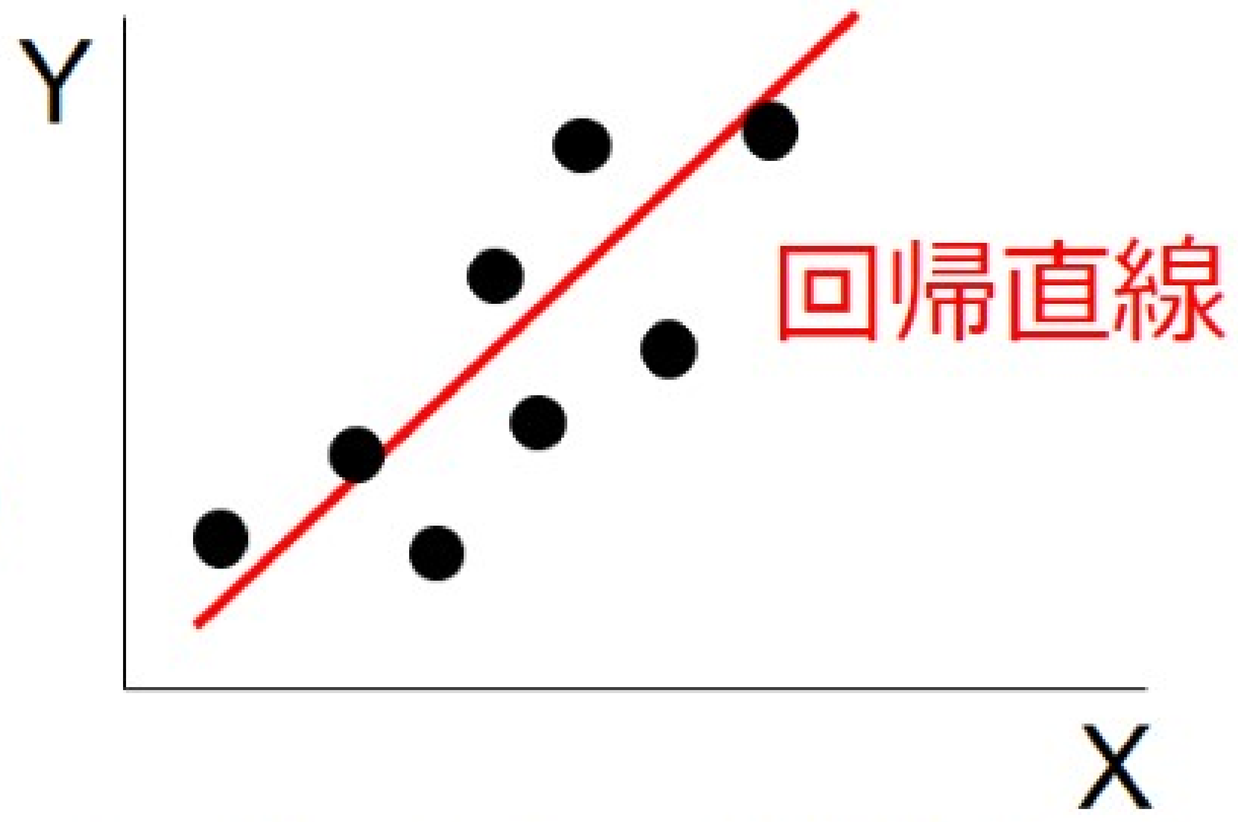
回帰

回帰直線とは



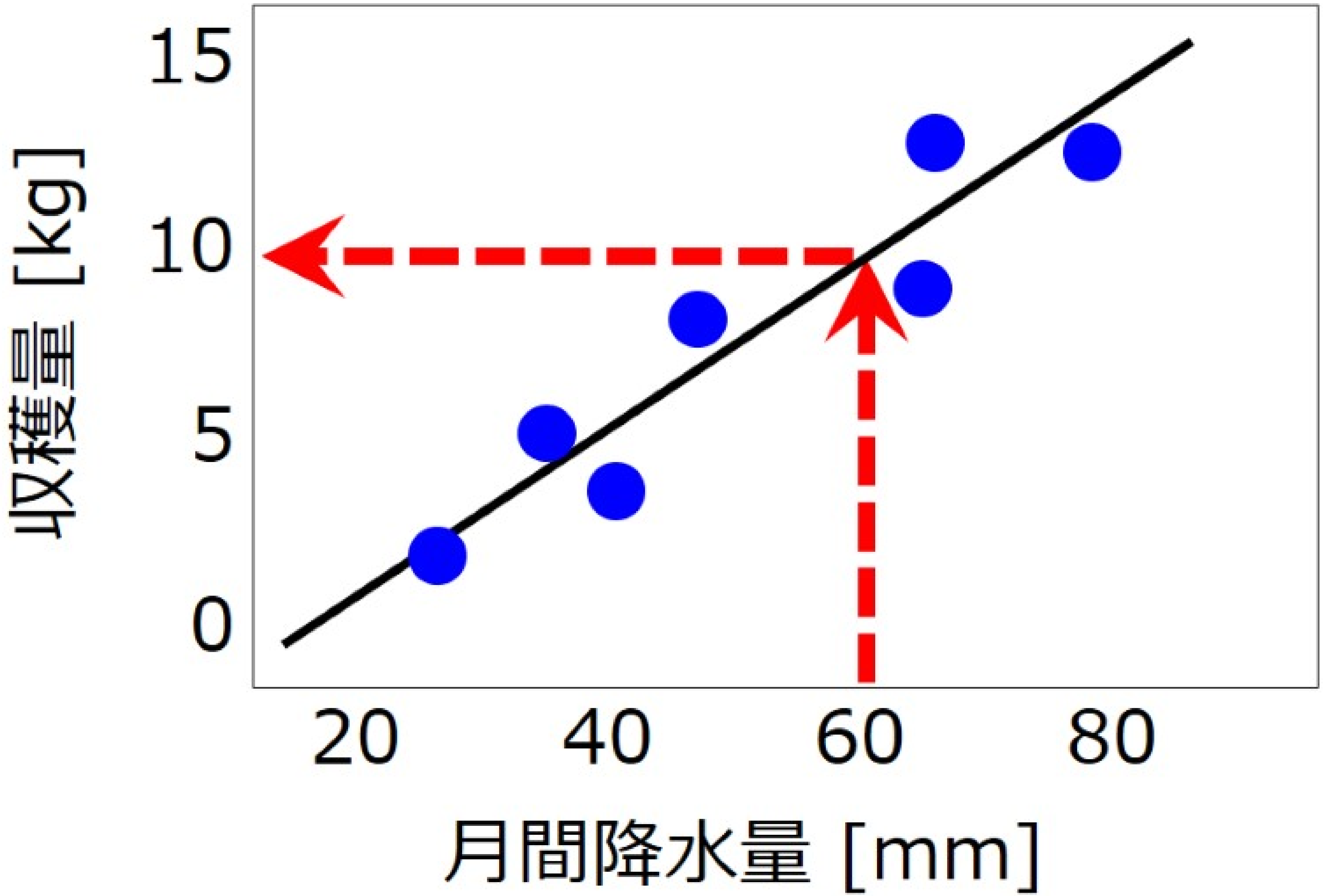
- これまで "なんとなく" 引いていた「データ間の関係を表す直線 (1次式)」を「**回帰直線**」と呼びます。
- X (たとえば月間降雨量) を「**説明変数**」、Y (たとえば収穫量) を「**目的変数**」と呼びます。回帰直線を引くということは、説明変数から目的変数をできるだけうまく推定するための関係式(回帰式) を求めていることになります。
  - $Y = a_0 + a_1 X_1$　説明変数が1つ:「線形単回帰」と呼ぶ
- 回帰直線の「傾き」を表す a_1 は「説明変数 X_1 が 1 増えたときに目的変数 Y がいくつ増えるか」を表します。

- 説明変数Xは1次元でなくてもかまいません。たとえば降雨量X_1，日照時間X_2，平均気温X_3を説明変数にして収穫量Yとの間の関係式を求めることもできます(次回に詳しく学びます)。
  - $Y = a_0 + a_1 X_1 + a_2 X_2 + a_3 X_3$

説明変数が2つ以上:「線形重回帰」と呼ぶ

目的変数の値の予測

- X(説明変数)とY(目的変数)の回帰式 (つまり a_0, a_1, ..., これらを偏回帰係数と呼ぶ) を求めることができれば、説明変数の値から目的変数を「予測」できます。

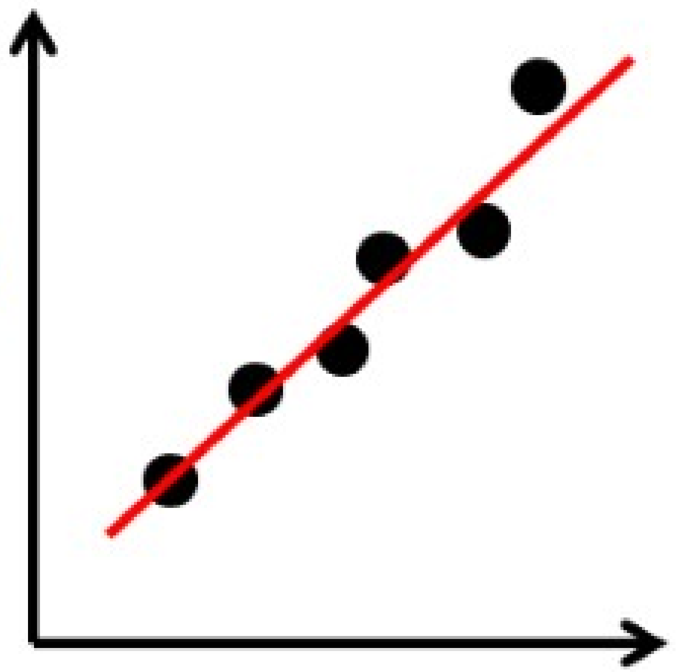


たとえば月間降水量が 60 mmであれば、収穫量は (多少ずれはあるにせよ) およそ 10 kgくらいではないか、という予測ができる。

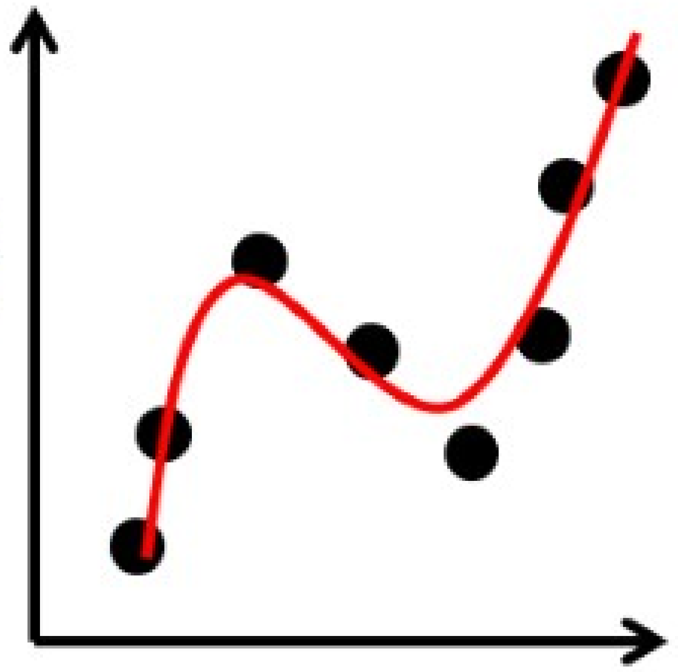
回帰分析

- このように、説明変数Xと目的変数Yとの間に、Y=f(X)という関係式 (数理モデル) を当てはめてデータを解析することを、回帰分析と呼びます
  - Xが1次元なら「単回帰」、2次元以上なら「重回帰」
  - 関数 f(X) が $Y=a_0 + a_1 * X$ の形式で表せる 1次式 (線形関数, 図形的には直線) なら「線形回帰」、非線形関数なら「非線形回帰」

線形回帰



非線形回帰



回帰分析の応用

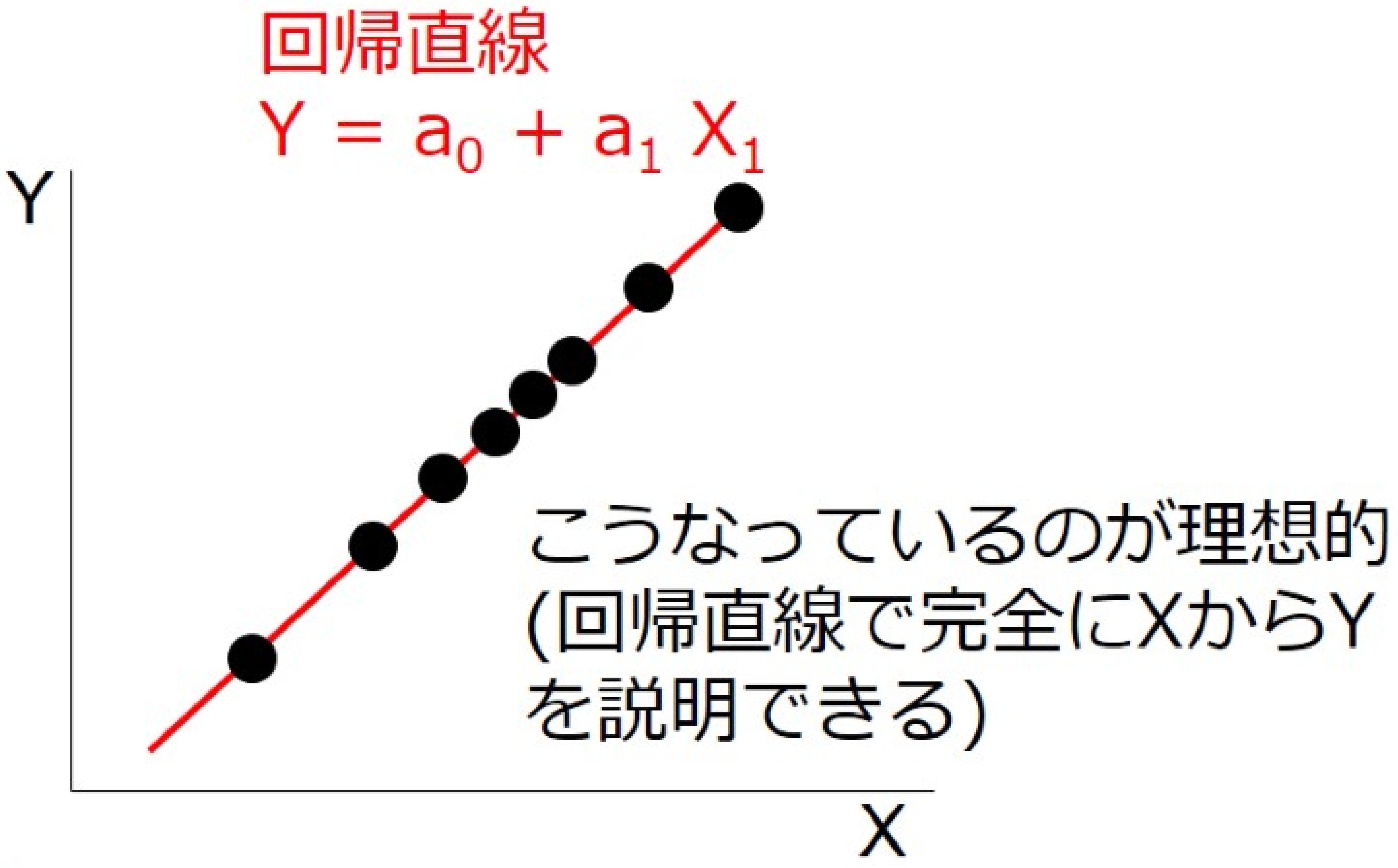
- 回帰分析は、ビジネス、科学、社会などあらゆる分野で非常によく用いられるデータ解析手法の1つです。
- 例: ある店舗の「1ヶ月売上高」の値を、「駅からの距離」などの店舗の属性を説明変数として回帰分析。

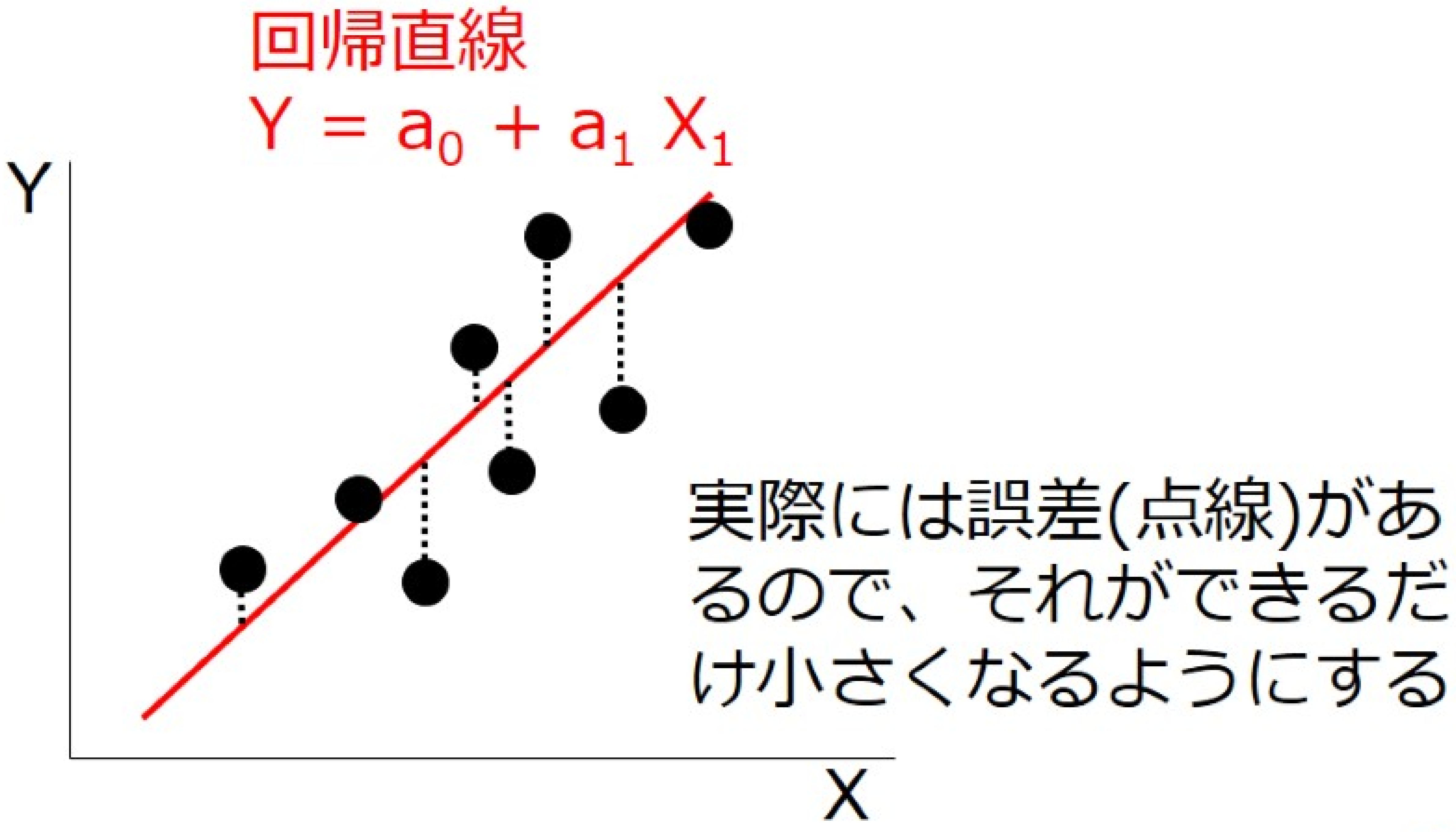
店舗ID	駅からの距離(km)	面積(m^2)	品数	店長年齢	1ヶ月売上高(千円)
1	4.0	35	124	58	942
2	1.7	24	82	42	760
3	0.3	20	76	38	425
…					

分析結果として、**面積や品数などの要因**は、それぞれ1ヶ月売上高にどのくらいの影響を与えているのか？を見積もることができます。
→ 売上高を伸ばすための方策は何が効果的？などの戦略立案ができる、
新たに出店する店舗のデータから、その店舗の売上高が予測できる、…

偏回帰係数の求め方 (線形単回帰)

- 回帰直線と目的変数（y軸方向）のずれ（下図の点線）ができるだけ小さくなるように、回帰直線の係数a_0 , a_1 , … を求めます。
- ずれは、正方向、負方向どちらもありうるので、「ずれの2乗和」が最小になるようにします (最小2乗法)。





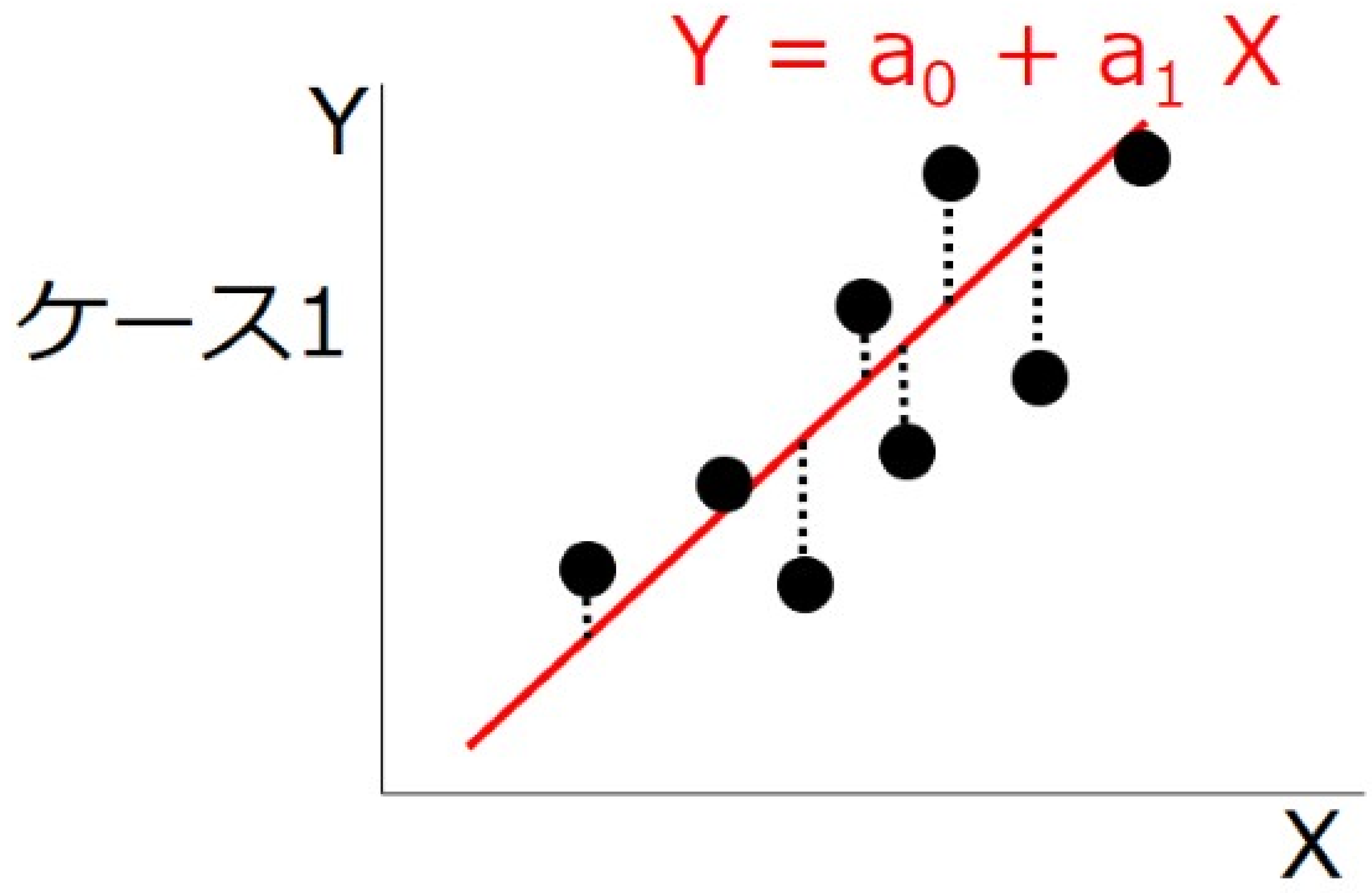
偏回帰係数の求め方 (線形単回帰)

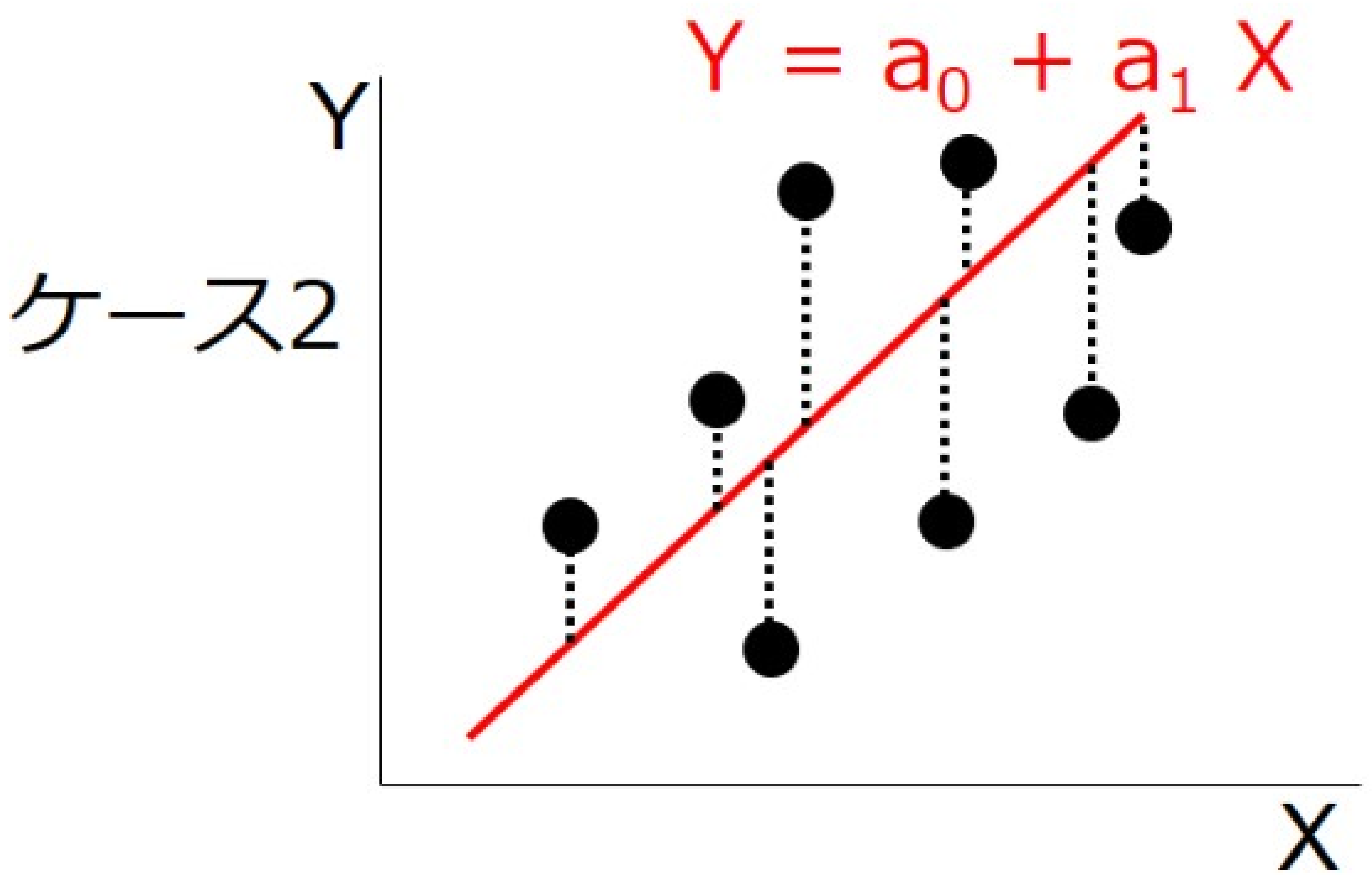
(発展) x が1次元の場合は、数学的にx, x^2, y, y^2, xy の和から、ずれの2乗和を最小にする a_0, a_1を求めることができる

$$a_0 = \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i{}^2\right) - \left(\sum_{i=1}^{n} x_i y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n\left(\sum_{i=1}^{n} x_i{}^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$a_1 = \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right) - n\left(\sum_{i=1}^{n} x_i y_i\right)}{\left(\sum_{i=1}^{n} x_i\right)^2 - n\left(\sum_{i=1}^{n} x_i{}^2\right)}$$

式の当てはまりはどのくらい？



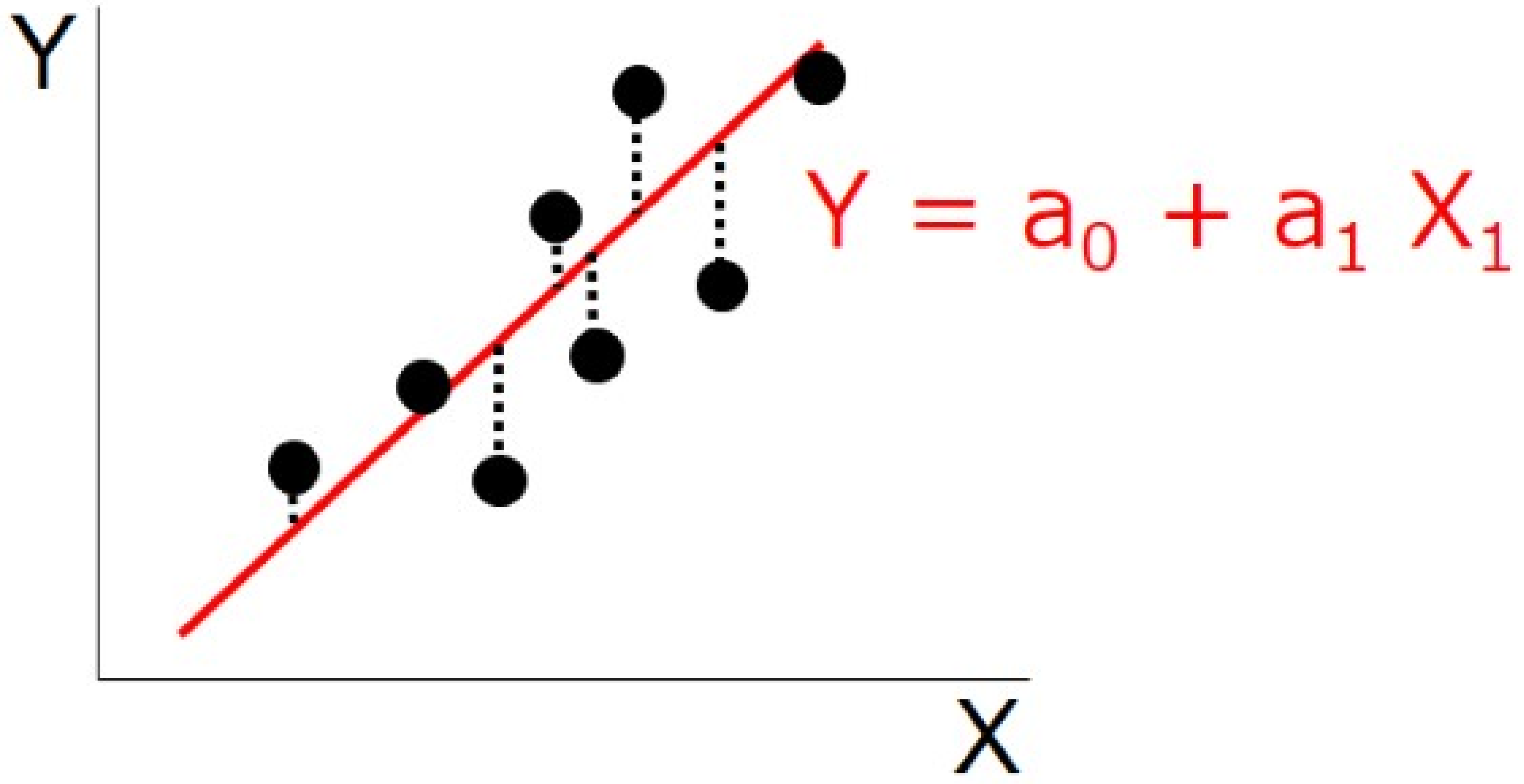


a_0, a_1 が同じでも、XとYの関係性が強く(相関係数の絶対値が大)、XからYを精度よく求められる場合(ケース1)と、 XとYの関係性が弱く(相関係数の絶対値が小)、Xから見積もったYと真のYの間の誤差が大きい場合(ケース2)があります。

➡ XとYの関係式(回帰直線)の当てはまりの程度(=YがXでどのくらい精度よく表せるか)を数値化することが必要

式の当てはまりはどのくらい？

- 目的変数Yが説明変数Xでどのくらい表せるかを示す指標として、平均2乗誤差 (MSE, Mean Squared Error) や決定係数R^2があります。
- 平均2乗誤差(MSE): 各点における回帰直線とYのずれ（図の点線の長さ）の2乗の平均値。各点の回帰直線からのずれが大きいほどMSEは大きくなります。
- 決定係数R^2: 1-(MSE/Yの分散)。もともとのYのばらつき (分散) を考慮した指標。全点が完全に直線上にあればMSE=0なのでR^2=1 (決定係数最大)。回帰直線からのずれが大きいほどR^2は小さくなります。



あくまで目安だが、R^2 が0.7以上であれば当てはまりはよく、0.9以上であれば非常によい当てはまり、逆に0.5未満であれば当てはまりは悪い、と一般的にいわれている

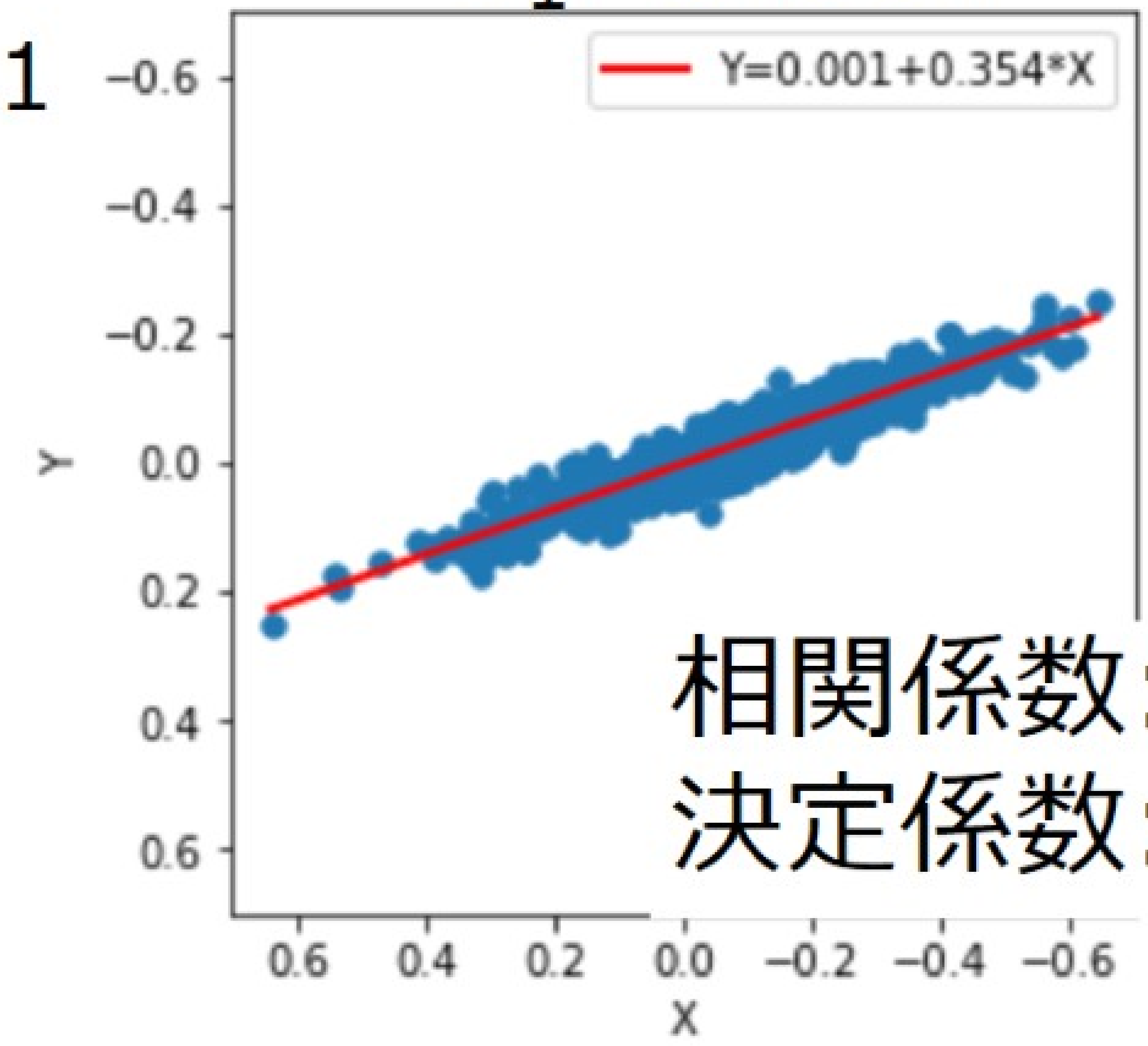
決定係数について

- X, Yの間の相関係数の2乗は、最小2乗法によるX, Yの回帰直線 (原点を通らなくてもよい) の決定係数と一致します。
- 説明変数の数が多くなると決定係数は1に近づくという性質があるので、説明変数が異なるモデル同士の当てはまりの比較には不適です。このため、説明変数の数で補正した「自由度調整済み決定係数」もよく用いられます。
  - 決定係数 = 1 – (MSE/Yの分散)
  - 自由度調整済み決定係数 = 1 – (MSE/Yの分散) * (N – 1) / (N – k – 1)
  - N: データ数、k: 説明変数の数
- なお、決定係数の定義は複数存在するので、各ソフトウェアで定義を確認することが望ましい。ここで説明しているのは、本講義で使用する statsmodelsライブラリのもので、もっとも標準的に使われている定義です。

注意! **線形単回帰**において、偏回帰係数 a1 は「XとYの関係性の強さ」を表す指標**ではない** $Y = a_0 + \mathbf{a_1} X_1$

a_1: 0.354

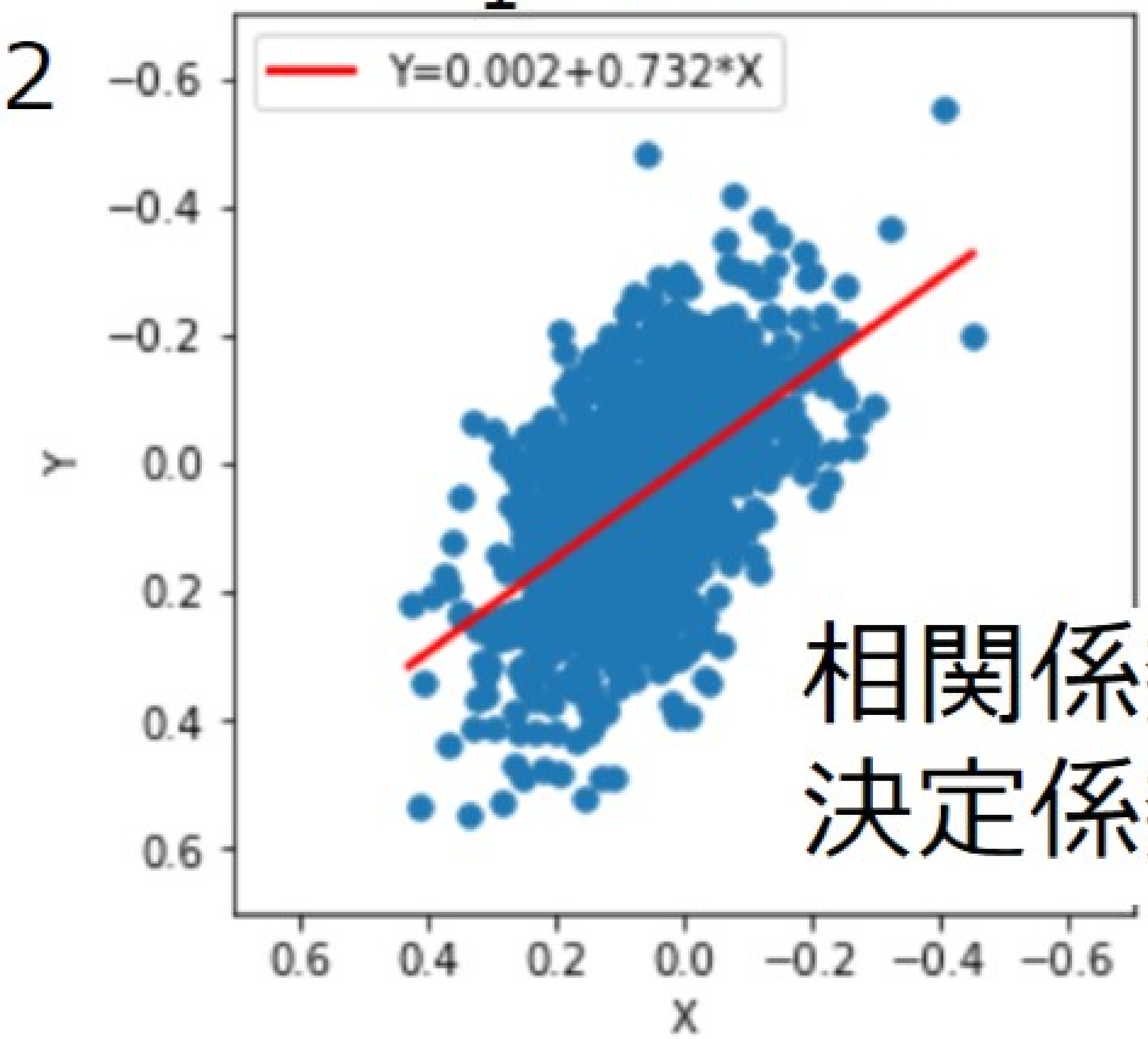
ケース1



相関係数: 0.942
決定係数: 0.887

a_1: 0.732

ケース2



相関係数: 0.545
決定係数: 0.297

a_1は回帰直線の傾きであり、X, Yの関係性の強さと直接関係はない。 X, Yの関係性の強さは相関係数や決定係数で測る。上図では、ケース1の方が a_1 は小さいが、相関係数や決定係数は大きく、 X, Yの関係性はケース2より強い。