情報科学
【AI・データサイエンス】

第7回
回帰分析と時系列分析

回帰分析
時系列分析

回帰分析

回帰とは？

「回って帰る」わけではないが，なぜか「回帰」と呼ばれる．
要は，与えられたデータに成り立つ傾向を見つけ出す方法

「テストの点数」を理解する

- 何が点数に影響するだろうか？

教室が快適だったから？

教え方が良い？

帰国子女だから？

沢山勉強したから？

読書量が多いから？

モデリング

- 観測（データ）を基に，現象を簡略化した「モデル（模型）」を作成すること
  - ここでは変数間の関係を数学的に表現するモデルを考える
- 例えば，テストの点数を以下のようにモデリングできる

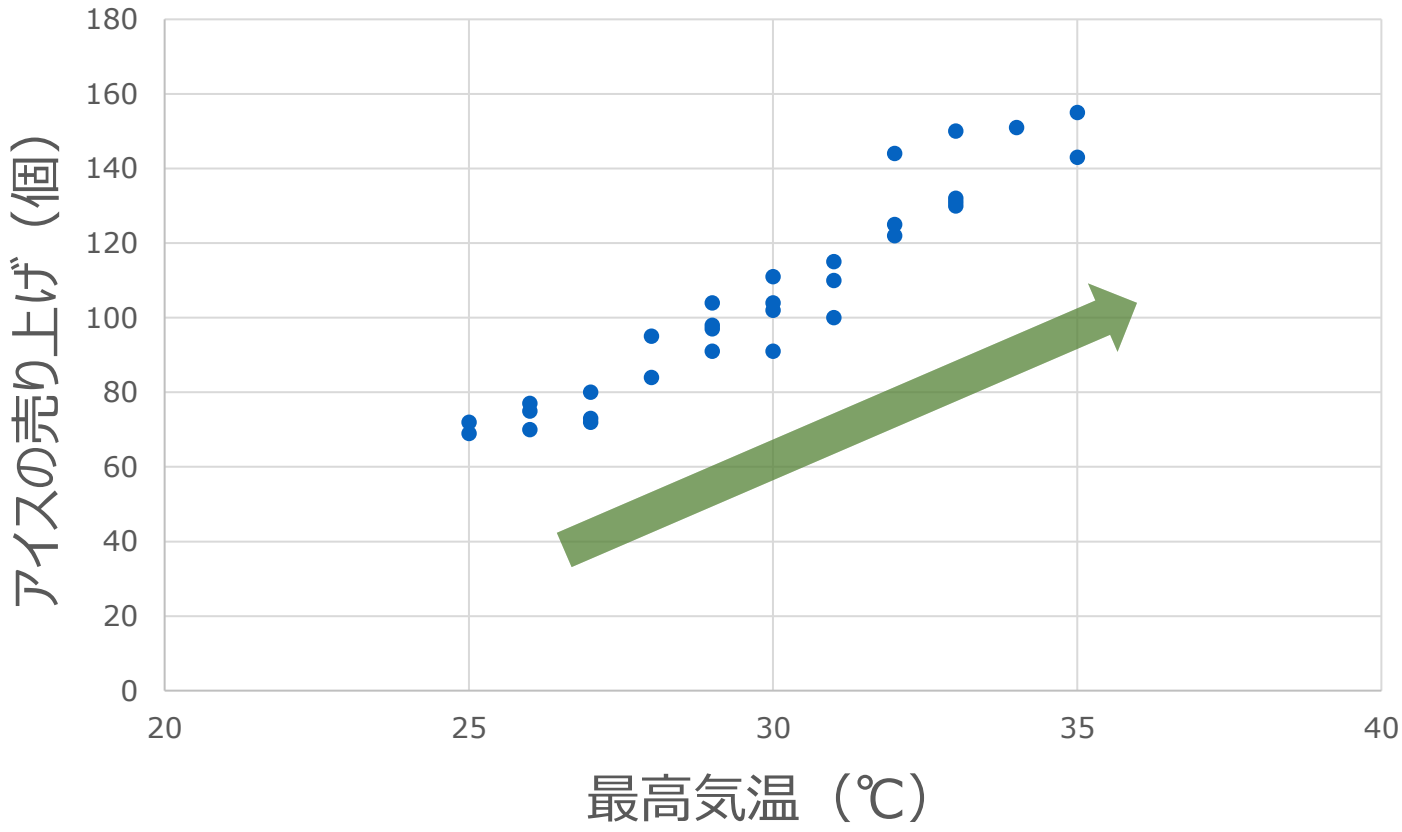
＝0.3×

＋0.7×

【テストの点数は，教え方3割，家での勉強7割】というモデル

- 変数：テストの点数，先生の授業経験年数，自宅での勉強時間
- ここでは特に線形回帰モデルの枠組みでモデリングしている
- モデルの正しさは別途要確認！

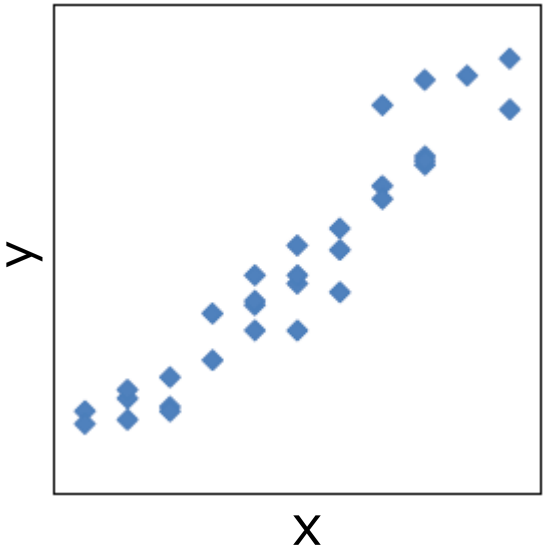
ほかの例：最高気温とアイスの売り上げの関係



最高気温が高くなると，アイスの売り上げも伸びる傾向

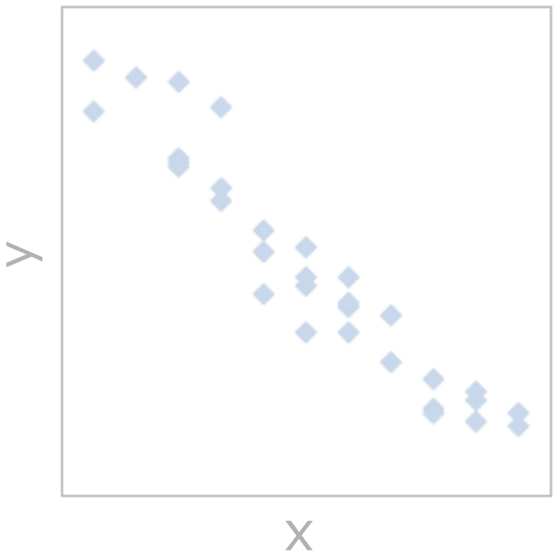
参考：相関分析を使うと…

- 正の相関



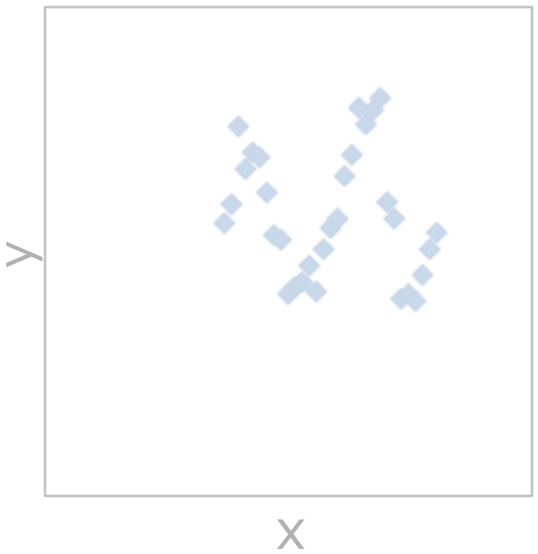
xが増加するとyも増加する

- 負の相関

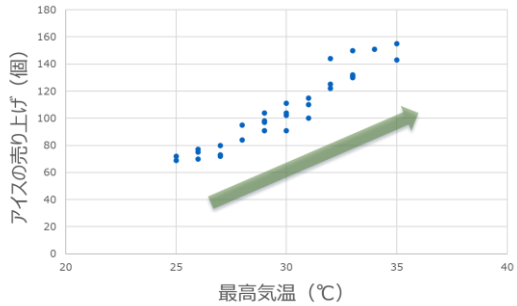


xが増加するとyが減少する

- 無相関



どちらにも当てはまらない



←「(ある程度強い)正の相関を持つ」
ぐらいしか分からない

回帰分析だと

- 変数（データ）間の関係式が分かる

日付	最高気温（℃）	アイスの売り上げ（個）
2014/07/01	30	102
2014/07/02	31	110
2014/07/03	35	143
2014/07/04	32	125
2014/07/05	33	132
2014/07/06	33	130
2014/07/07	31	115
2014/07/08	29	97
2014/07/09	28	95
2014/07/10	25	72
2014/07/11	26	75
2014/07/12	26	77
2014/07/13	28	84
2014/07/14	27	73
2014/07/15	30	91
2014/07/16	31	100
2014/07/17	30	104
2014/07/18	32	122
2014/07/19	33	131
2014/07/20	34	151
2014/07/21	32	144

回帰式

（アイスの売り上げ）＝8.8×（最高気温）-158

目的変数
（従属変数）

説明変数
（独立変数）

66個くらいかな

25.5℃の時は？

回帰式を用いた予測

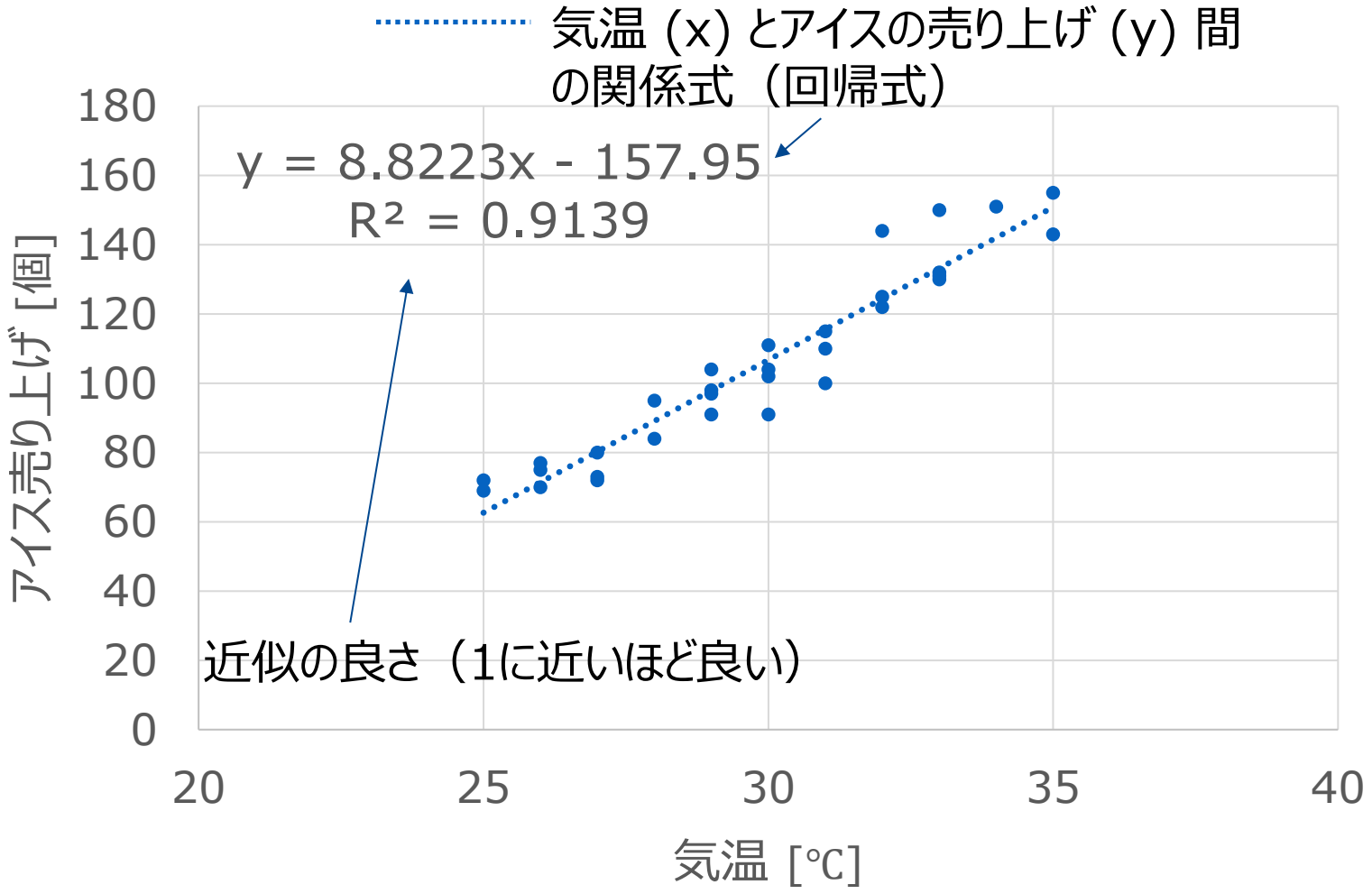
回帰分析とは

- データの属性の間の関係式を求める分析手法
  - 現象の理解や，未知の状況における予測に用いられる
  - データが時間的に独立である場合に使用
  - 目的変数を説明変数により記述する（回帰式を作る）

回帰式　（アイスの売り上げ）＝8.8×（最高気温）-158

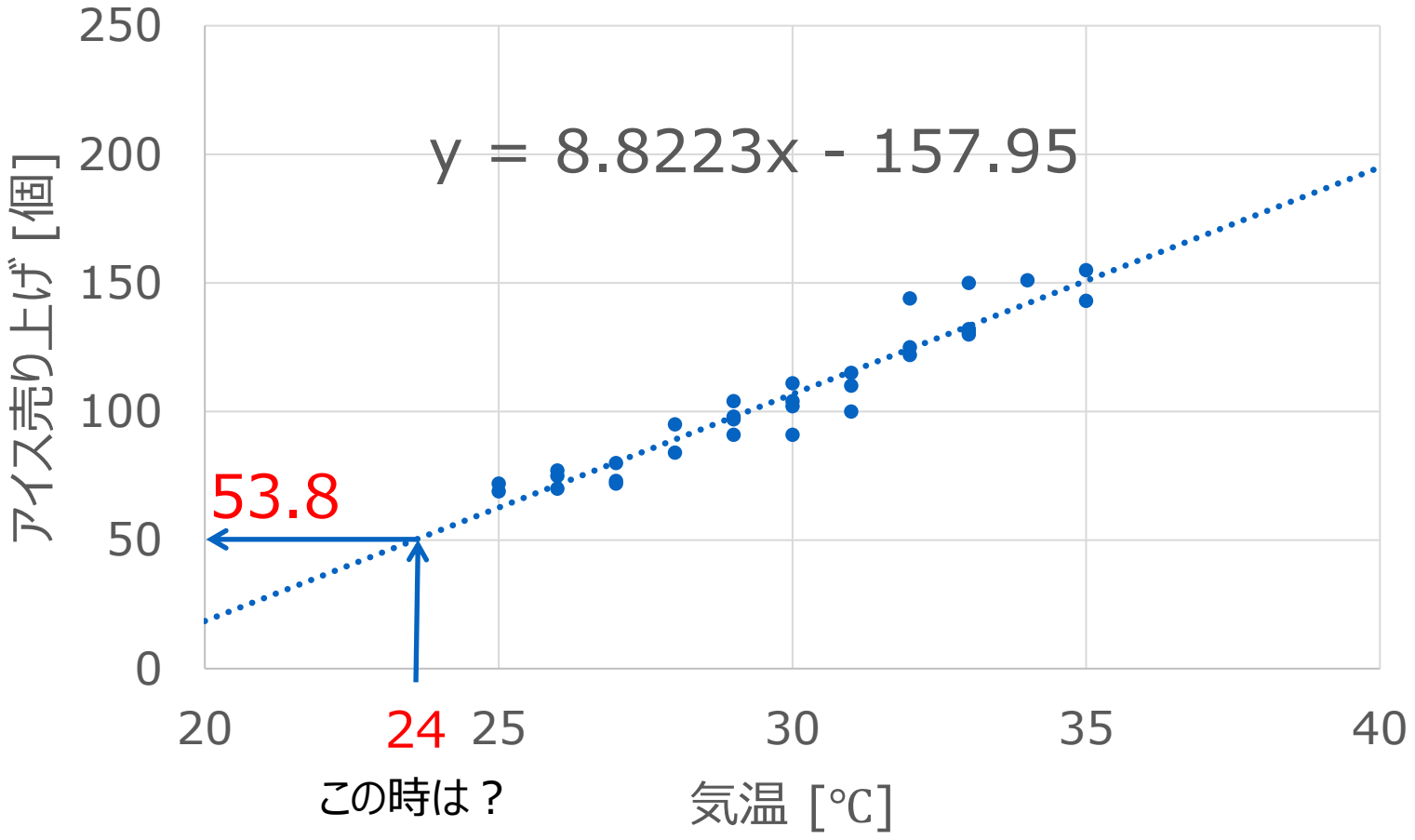
- 言葉の説明 (アイスの例で)
  - 1データ(点)は「1営業日」を表す
  - データの属性として「売り上げ」や「最高気温」がある
  - 今回はこれらの属性の内，売り上げを目的変数，気温を説明変数とした

回帰式を用いたモデルあてはめ



上記は線形モデル

回帰式を用いた予測：
未知の状況についても「これぐらいだろう」と予測がつく



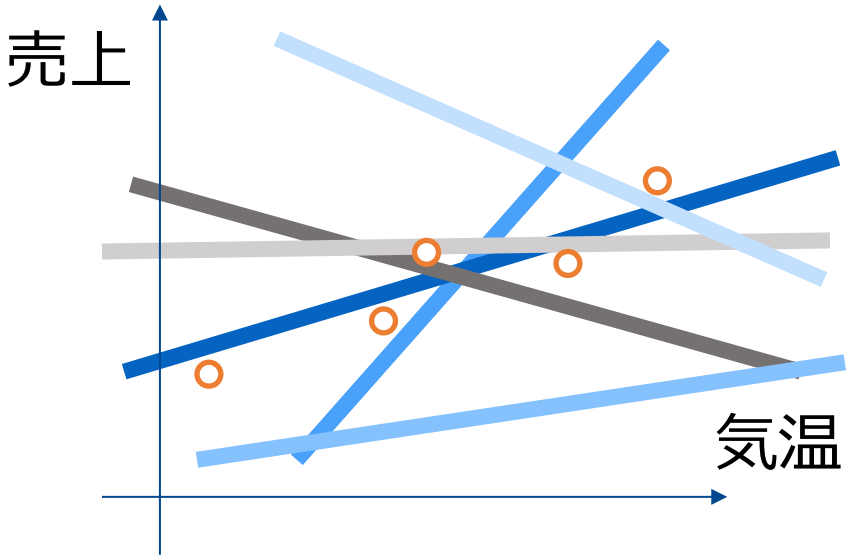
回帰の方法

「えいやっ」と線を引いてしまう

線形モデル

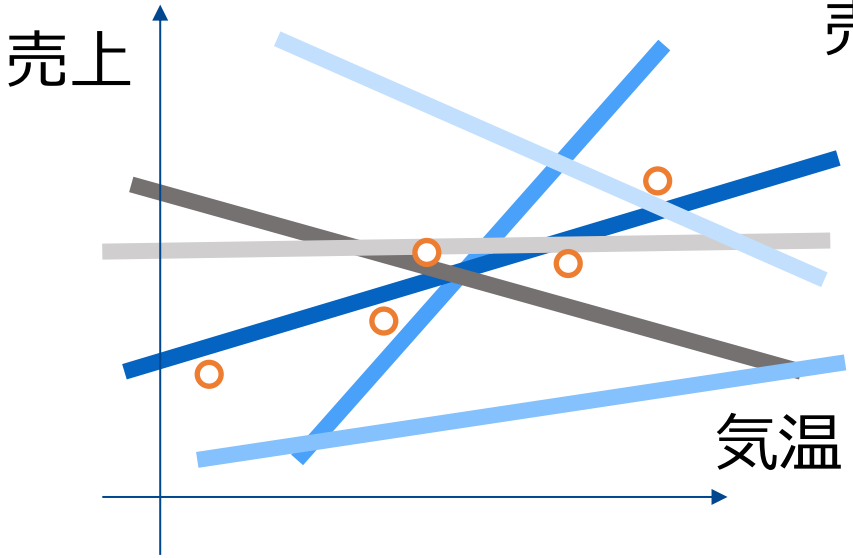
回帰式が$y = ax + b$の形のモデル

- aやbの具体値の組み合わせが1つのモデルに相当
- 直線がデータ点にあてはまるようなa, bの値をデータから計算により求める



モデルあてはめの方法

どういうあてはめがよい？



これが小さいほうがよい

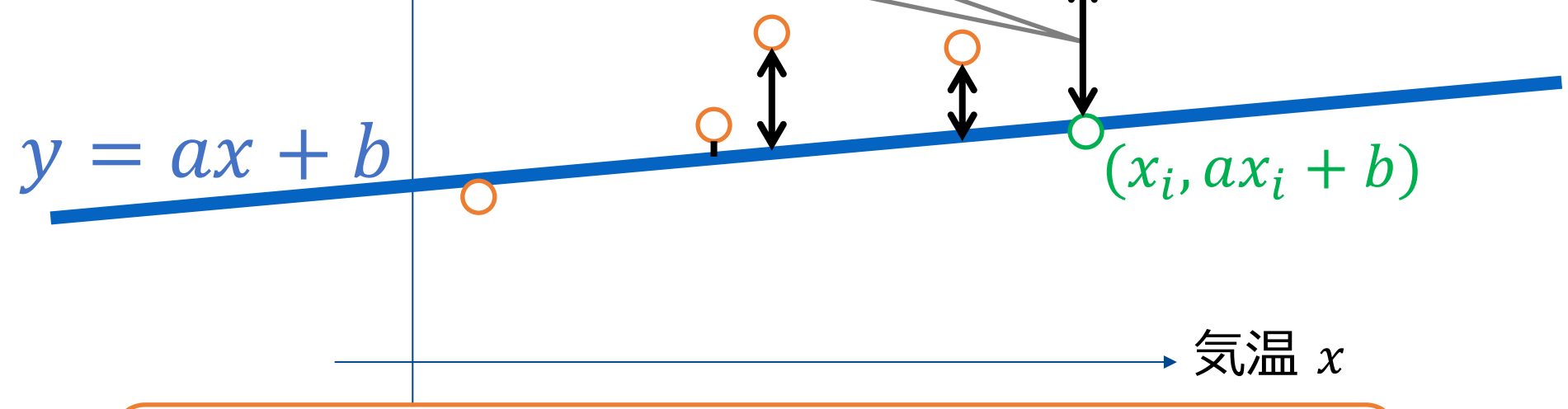
$\sum_i$ 第iデータのあてはめ誤差 → 最小化

最小二乗法

売上 y

二乗誤差で「あてはめ誤差」を定義

二乗誤差は$\left(y_i - (ax_i + b)\right)^2$



気温 x

aとbをいじって$\sum_i \left(y_i - (ax_i + b)\right)^2$を最小化

回帰分析の際の注意

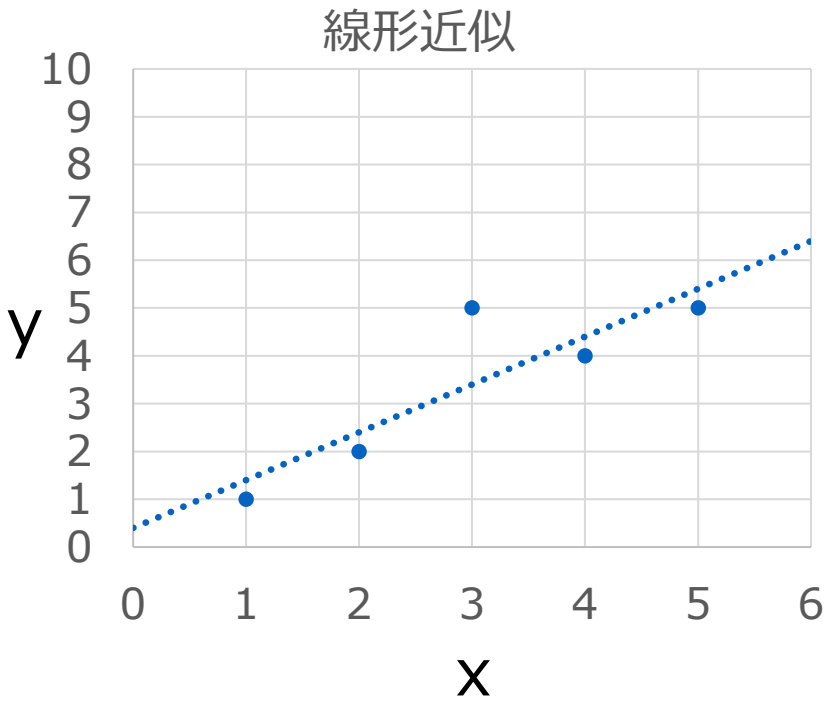
「過ぎたるは及ばざるがごとし」

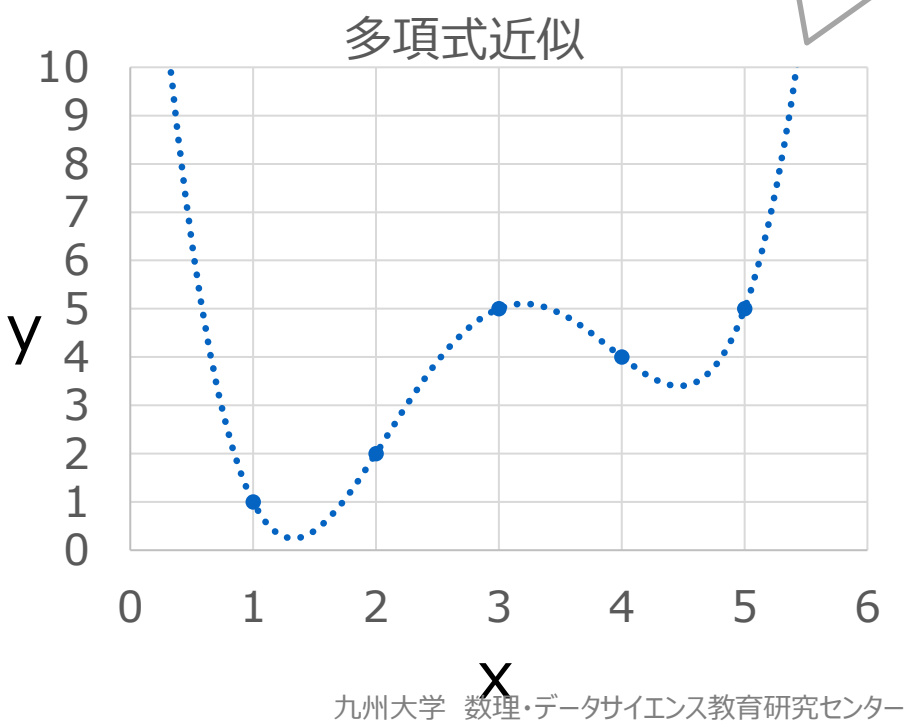
「線形近似」と「より複雑な近似」（多項式近似）のどちらが良いのか？

例：元データ

x	y
1	1
2	2
3	5
4	4
5	5

外れ値（計測誤差や例外的なデータ）

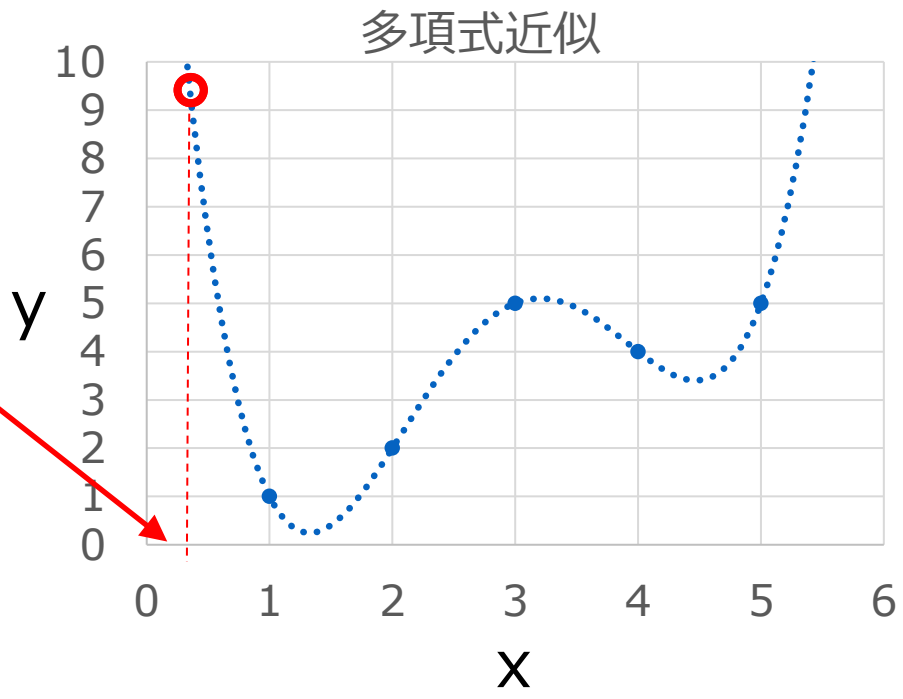




誤差はゼロだけど，それでいいの？：オーバーフィッテイング

観測データに対しては
良くあてはまっている

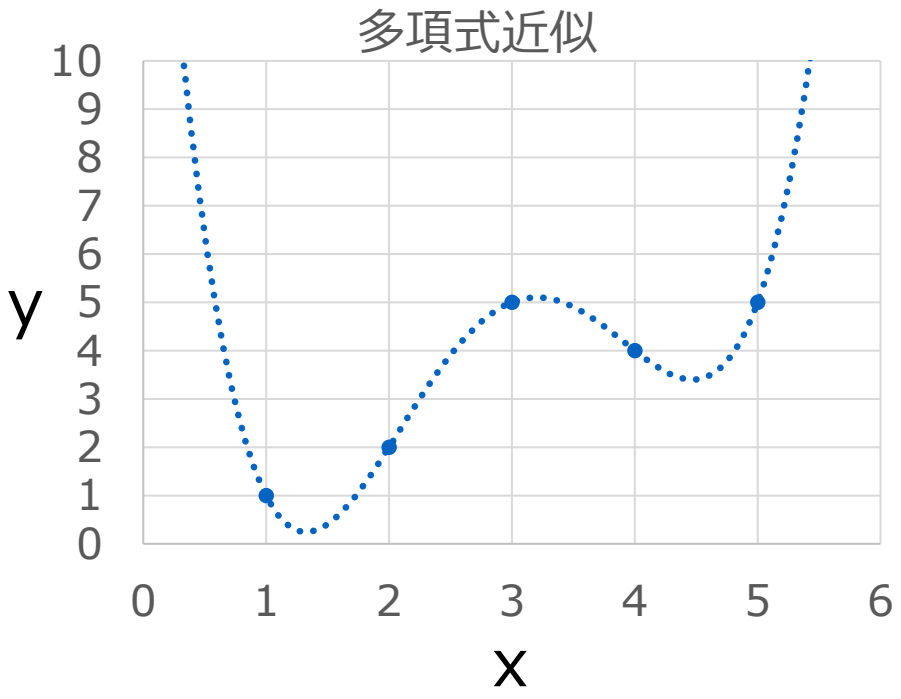
一方で，観測されていない場合(xが1，2，3，4，5でない場合)に関しては使い物にならないことが多い= 汎化能力が低い



これをオーバーフィッティング(過剰適合)しているという

汎化能力とは

- 回帰曲線を求めるときに**データになかったx**(＝未知の場合)についても妥当な予測結果が得られるかどうか



これ↑は汎化能力がない例

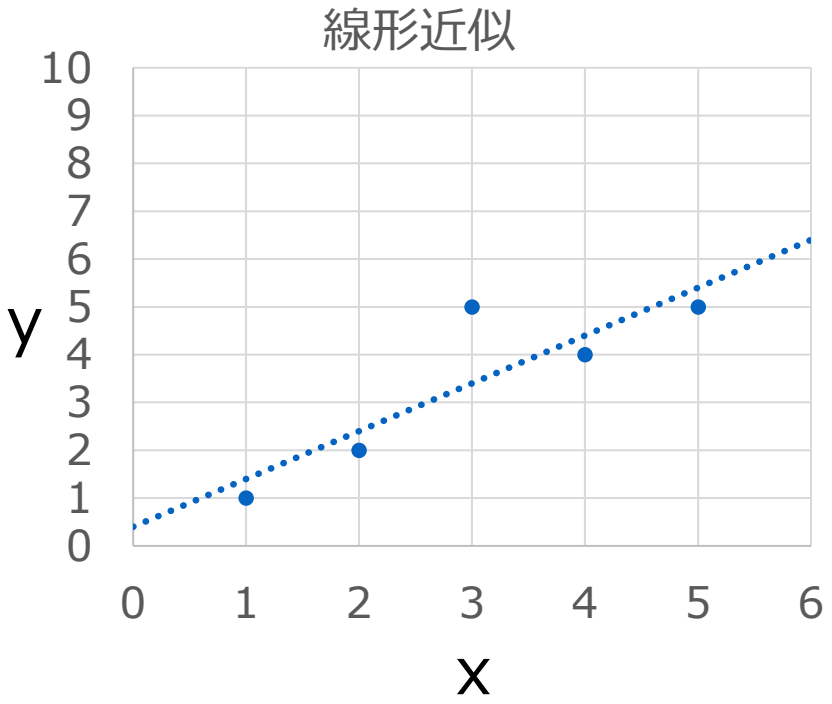
余談：日常にもあるオーバーフィッティング

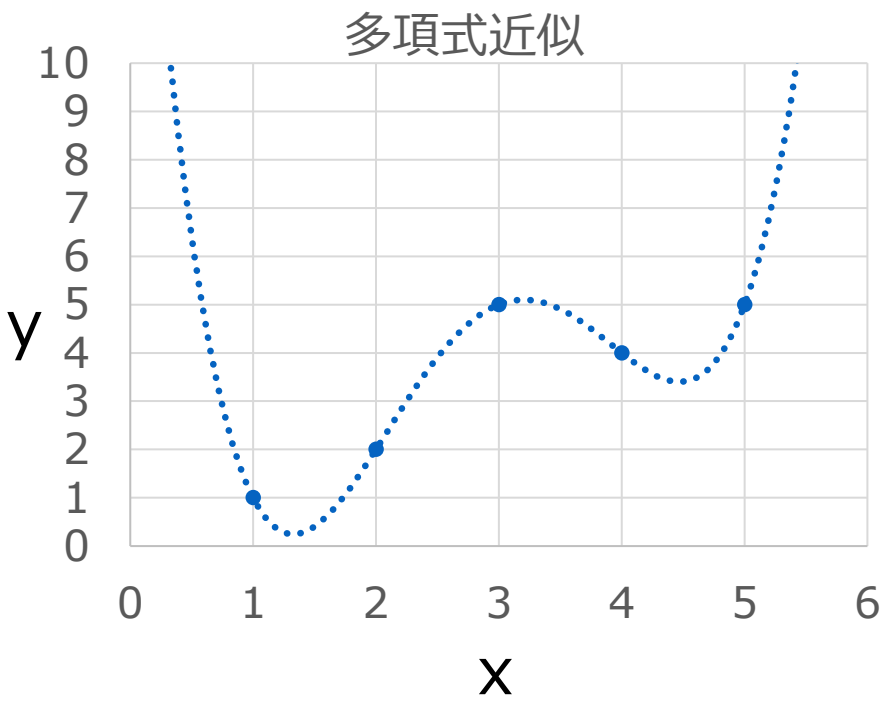
- 自宅の問題集は完璧だけど，そればっかりやりすぎて，他の問題には全く応用が利かない

- ある「声の大きな人（実力者）」の意見を聞きすぎて，施策を決めてしまったところ，ほかの多くの人から文句が出た

- 自分のパートナーのことが素敵に見えすぎて，ほかの人は全員ダメダメに見える
  - それはそれで悪くない？

線形近似と多項式近似のどちらが良いのか？





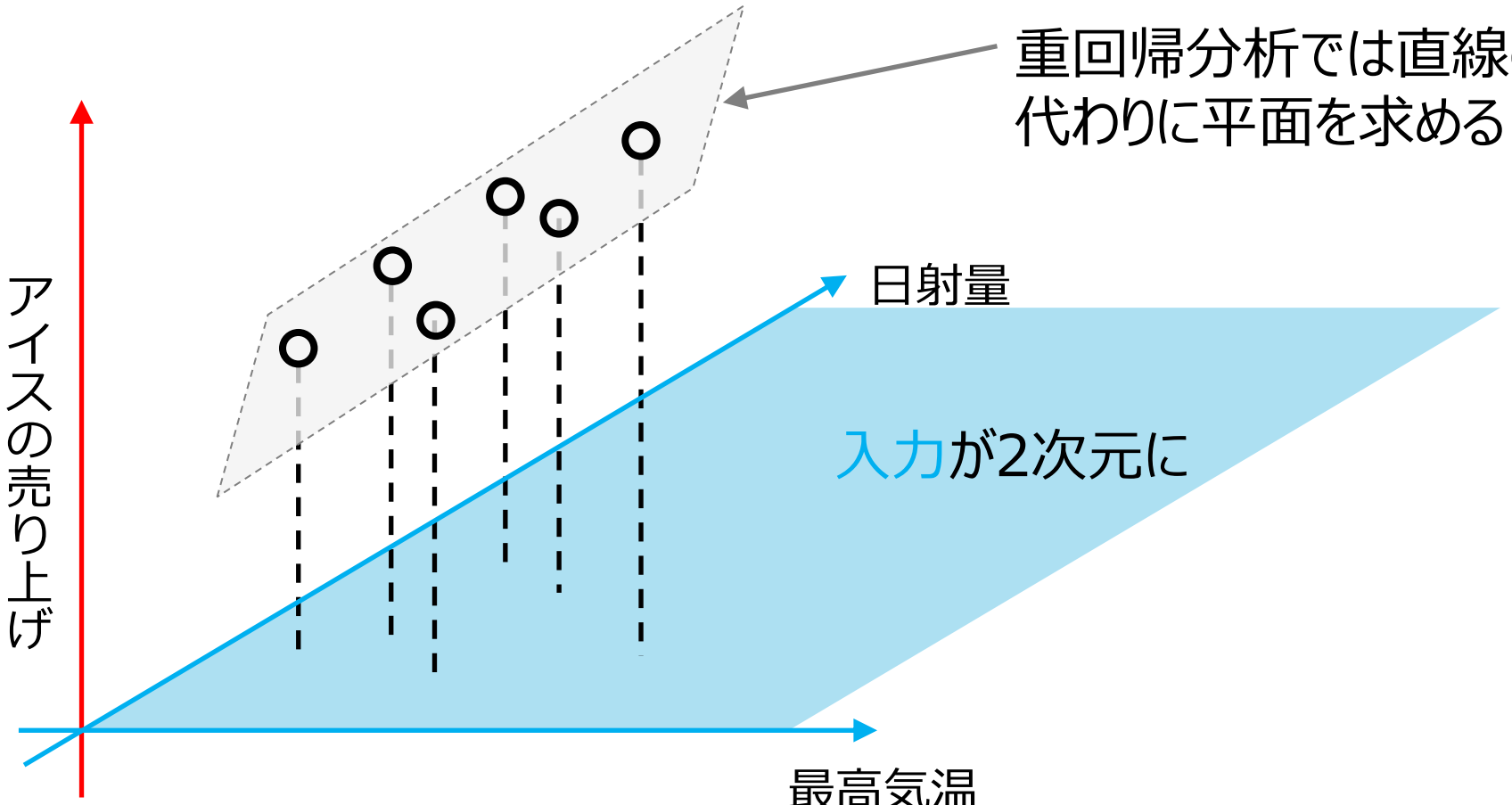
（良） 事前データ数が少なくてもひどいオーバーフィッティングは少ない
（悪） 事前に与えられたデータに対する誤差が大きい

（良） 事前に与えられたデータに対する誤差が小さい
（悪） 大量に事前データがないと回帰結果に汎化能力がない恐れ

参考：重回帰分析
（「重たい」わけではないです．「重」はMultipleの意味）

これまでの説明は単回帰分析

説明変数が2つ以上の場合を重回帰分析という



時系列分析

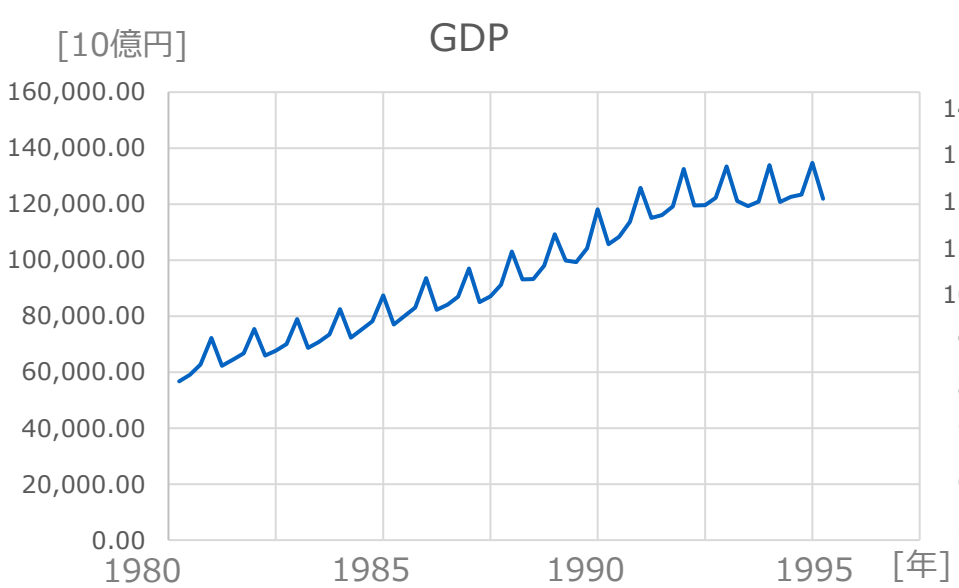
未来を知るための時系列「予測」を中心に

時系列データとは

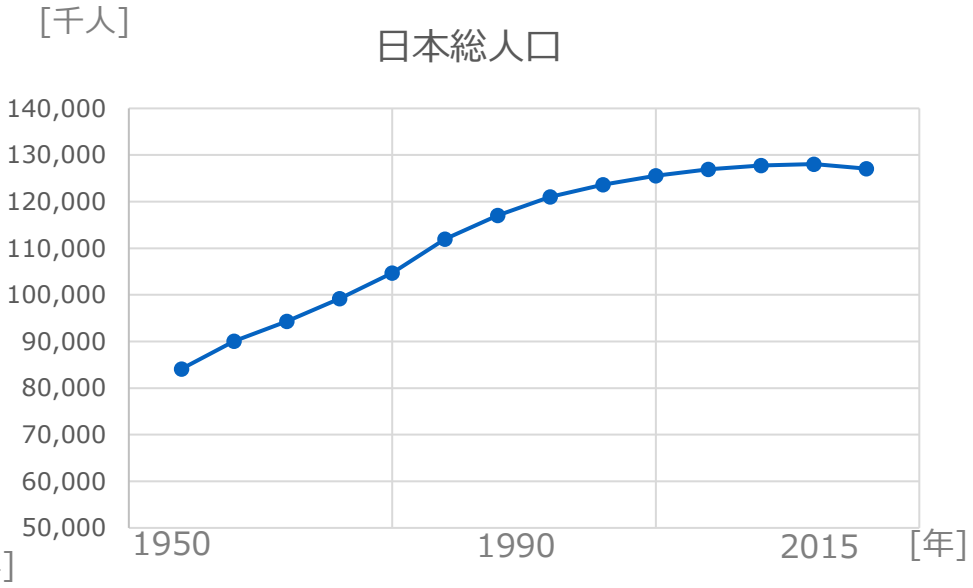
- 時間の推移とともに観測されるデータ

観測される順序に意味があることが大きな特徴

時系列データの例：

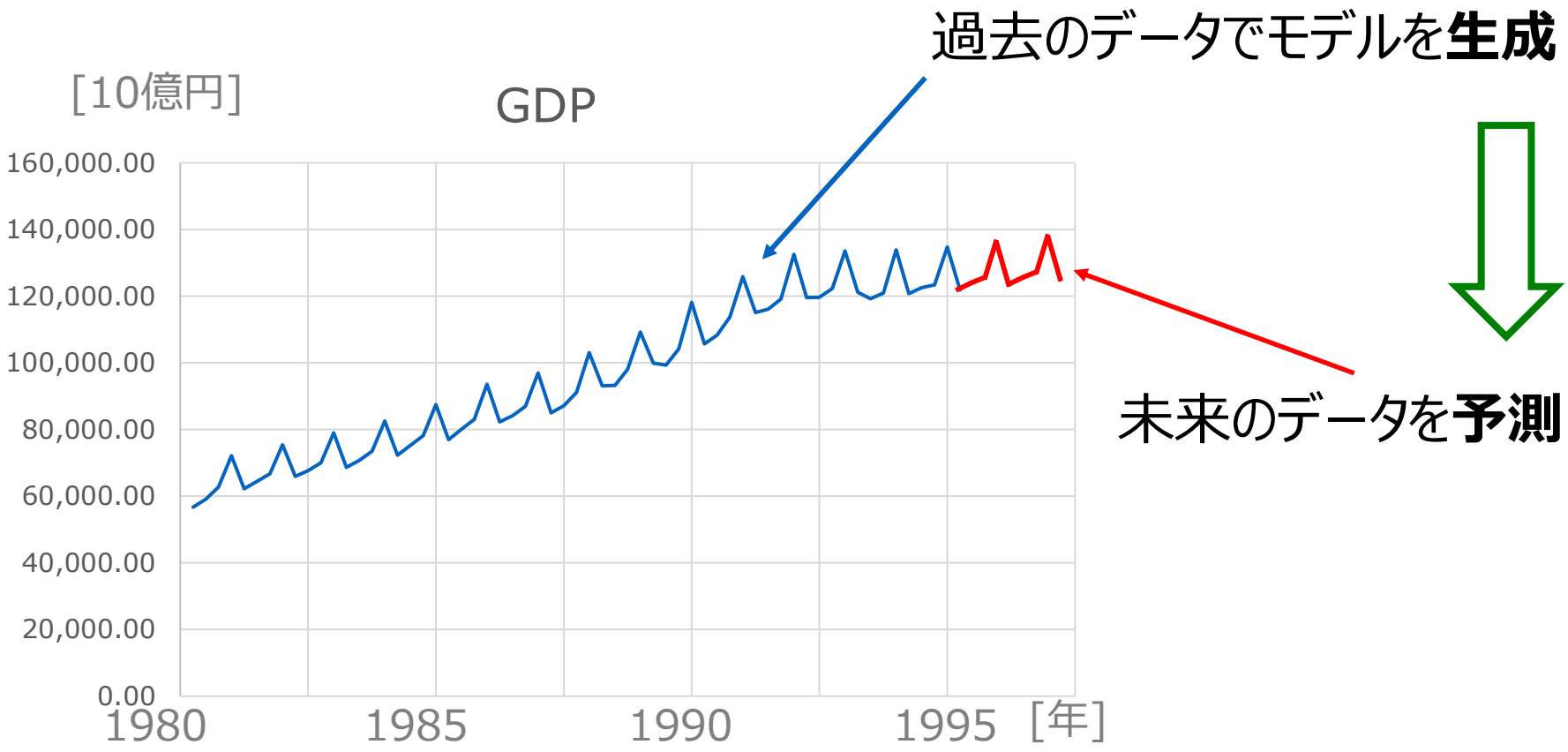


国内総生産（GDP）四半期ごと



日本総人口の推移

時系列分析の応用：予測



国内総生産（GDP）四半期ごと

時系列モデルを用いた予測の流れ

1. 時系列データの取得

2. 時系列データの分析

   トレンドおよび季節成分を除いたランダム部分（定常時系列と呼ぶ）を抽出

3. 分析結果から適切な時系列モデルの作成

4. 作成された時系列モデルを用いて未来のデータを予測

さまざまな時系列データのモデル

- 自己回帰モデル（ARモデル）
  時系列自身の過去の値を説明変数とする回帰
- 移動平均モデル（MAモデル）
- ARMA
  - AR・MAの両方を考慮
- ARIMA
  - データの時間差分に対してARMAを適用
- SARIMA
  - ARIMAに対してさらに周期的な変動を考慮

時系列モデルの推定方法

- 最小二乗法
  - モデルが説明できない部分（残差）の平方和が最小になるようにパラメータを設定する
- 最尤推定
  - 得られた観測値をモデルが最も実現しやすくなるようにパラメータを設定する

まとめ

- 回帰分析
  - データ間の関係式を求め予測に役立てる手法
    - 線形もでる
    - 単回帰
- 時系列分析
  - 仮説の検証や予測に役立てる方法
    - 時系列モデル

演習資料

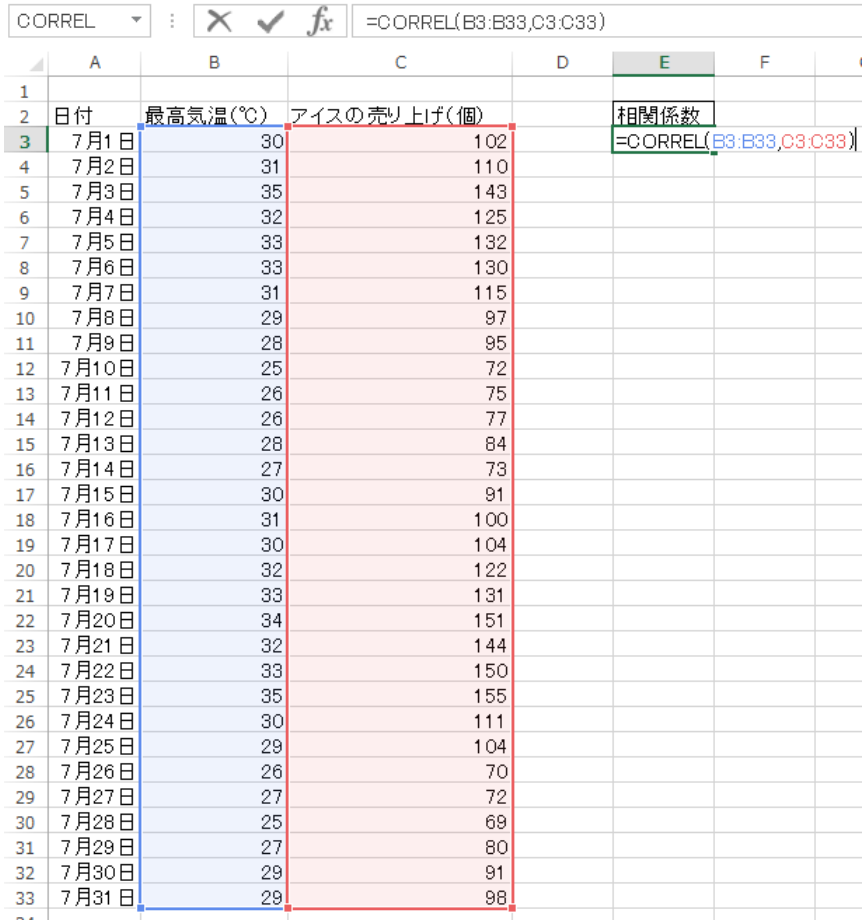
回帰・時系列

演習

- 国別のノーベル賞受賞者数と大学進学率について相関分析を行う
  - 散布図の作成
  - 相関係数の計算
  - 上記結果の基づく考察（相関はどうなのか）

- 身長と体重のデータについて回帰分析を行う
  - 散布図の作成（男女別）
  - 男女別に回帰式（線形）を求める
  - 求めた回帰式を用いて与えられた入力（身長）に対する体重の予測を行う

相関係数の計算（アイスの例）



	A	B	C	D	E	F
1						
2	日付	最高気温(℃)	アイスの売り上げ(個)		相関係数	
3	7月1日	30	102		=CORREL(B3:B33,C3:C33)	
4	7月2日	31	110			
5	7月3日	35	143			
6	7月4日	32	125			
7	7月5日	33	132			
8	7月6日	33	130			
9	7月7日	31	115			
10	7月8日	29	97			
11	7月9日	28	95			
12	7月10日	25	72			
13	7月11日	26	75			
14	7月12日	26	77			
15	7月13日	28	84			
16	7月14日	27	73			
17	7月15日	30	91			
18	7月16日	31	100			
19	7月17日	30	104			
20	7月18日	32	122			
21	7月19日	33	131			
22	7月20日	34	151			
23	7月21日	32	144			
24	7月22日	33	150			
25	7月23日	35	155			
26	7月24日	30	111			
27	7月25日	29	104			
28	7月26日	26	70			
29	7月27日	27	72			
30	7月28日	25	69			
31	7月29日	27	80			
32	7月30日	29	91			
33	7月31日	29	98			

Excelでは以下の関数を使う
CORREL(配列1, 配列2)

最高気温とアイスの売り上げの相関係数は 0.95
正の相関があるといえる

回帰直線の表示（アイスの例）

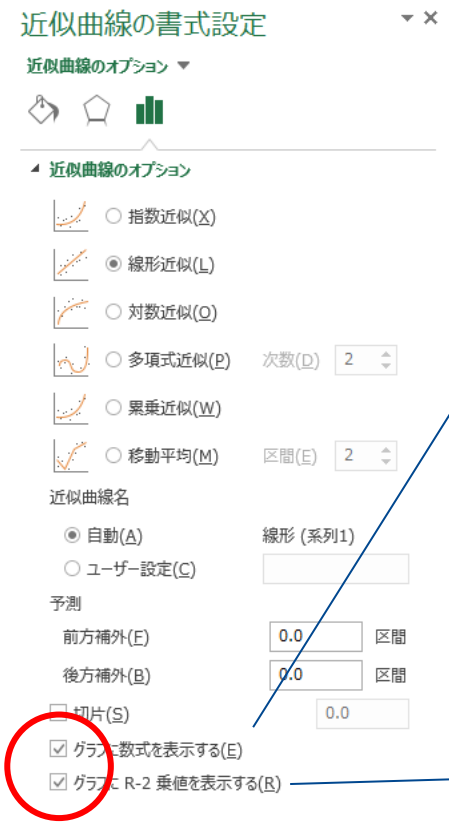
2種類のデータ間の関係式を求める分析手法
データの予測などに用いられる

例：「グラフ要素を追加」「近似曲線」「線形」を使うとデータを直線近似できる

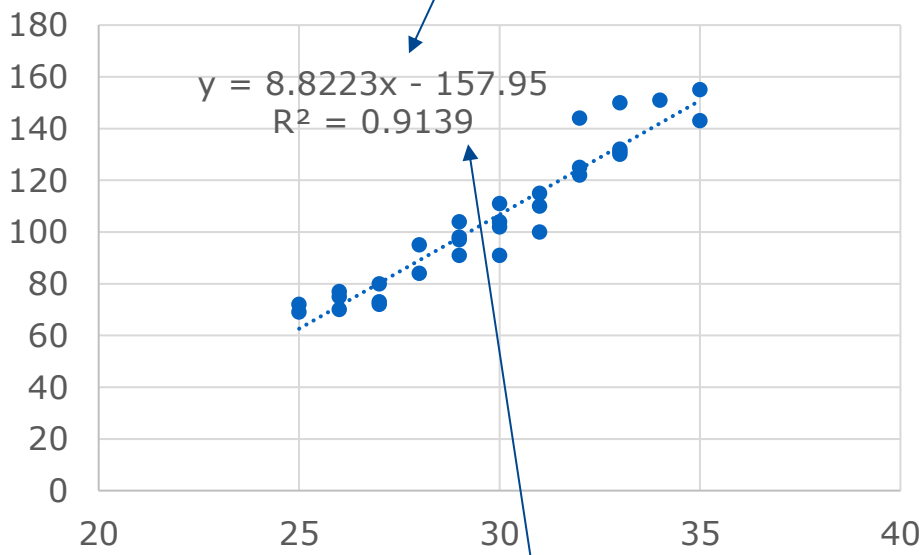
他にも複雑な曲線で近似可能

回帰式の算出（アイスの例）

「グラフ要素を追加」　「近似曲線」
「その他の近似曲線オプション」



気温（x）とアイスの売り上げ（y）間の関係式（回帰式）

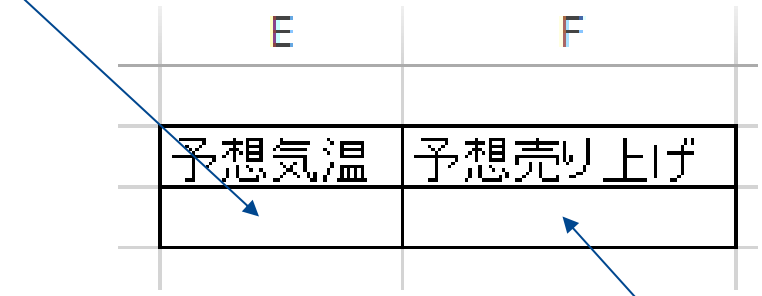


近似の度合い（1に近いほど良い近似）

回帰式を用いた予測（アイスの例）

ある気温 (E3) のときのアイス売り上げを予測する

E3のセルに任意の数値を入力



F3のセルに先ほど求めた回帰式「=8.8223*E3-157.95」を入力