



総務省統計局

社会人のためのデータサイエンス演習

第4週:ビジネスにおける予測と分析結果の報告

第1回:回帰分析による予測

講師名 矢島 安敏

講座内容

第1週

- データサイエンスとは

第2週

- 分析の概念と事例
ビジネス課題解決のためのデータ分析基礎(事例と手法)①

第3週

- 分析的具体的手法
ビジネス課題解決のためのデータ分析基礎(事例と手法)②

第4週

- ビジネスにおける予測と分析結果の報告
ビジネス課題解決のためのデータ分析基礎(事例と手法)③

第5週

- ビジネスでデータサイエンスを実現するために

第4週の内容紹介

第1回

- 回帰分析による予測

第2回

- モデル評価と予実評価

第3回

- 分析結果の報告（記述/可視化方法）

第4回

- 分析結果の報告（解釈の注意点）

第5回

- 予測・分類等代表的手法と活用場面

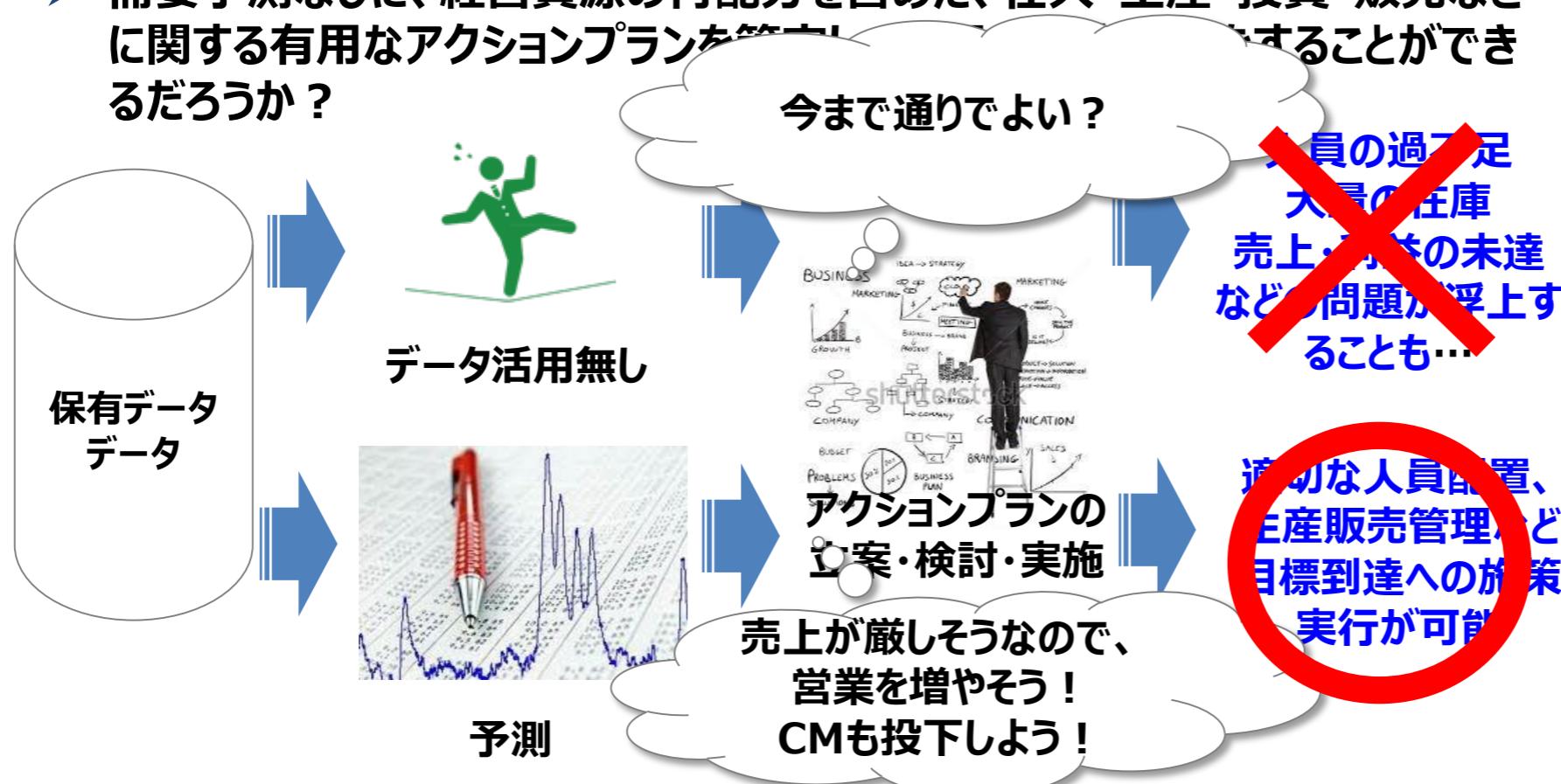
第6回

- ビジネスシーンにおける「統計的検定」とその活用例

予測とは

- ビジネスでの予測とは、将来の経営計画、事業計画などを策定し、有用な行動する上で欠かせない作業である

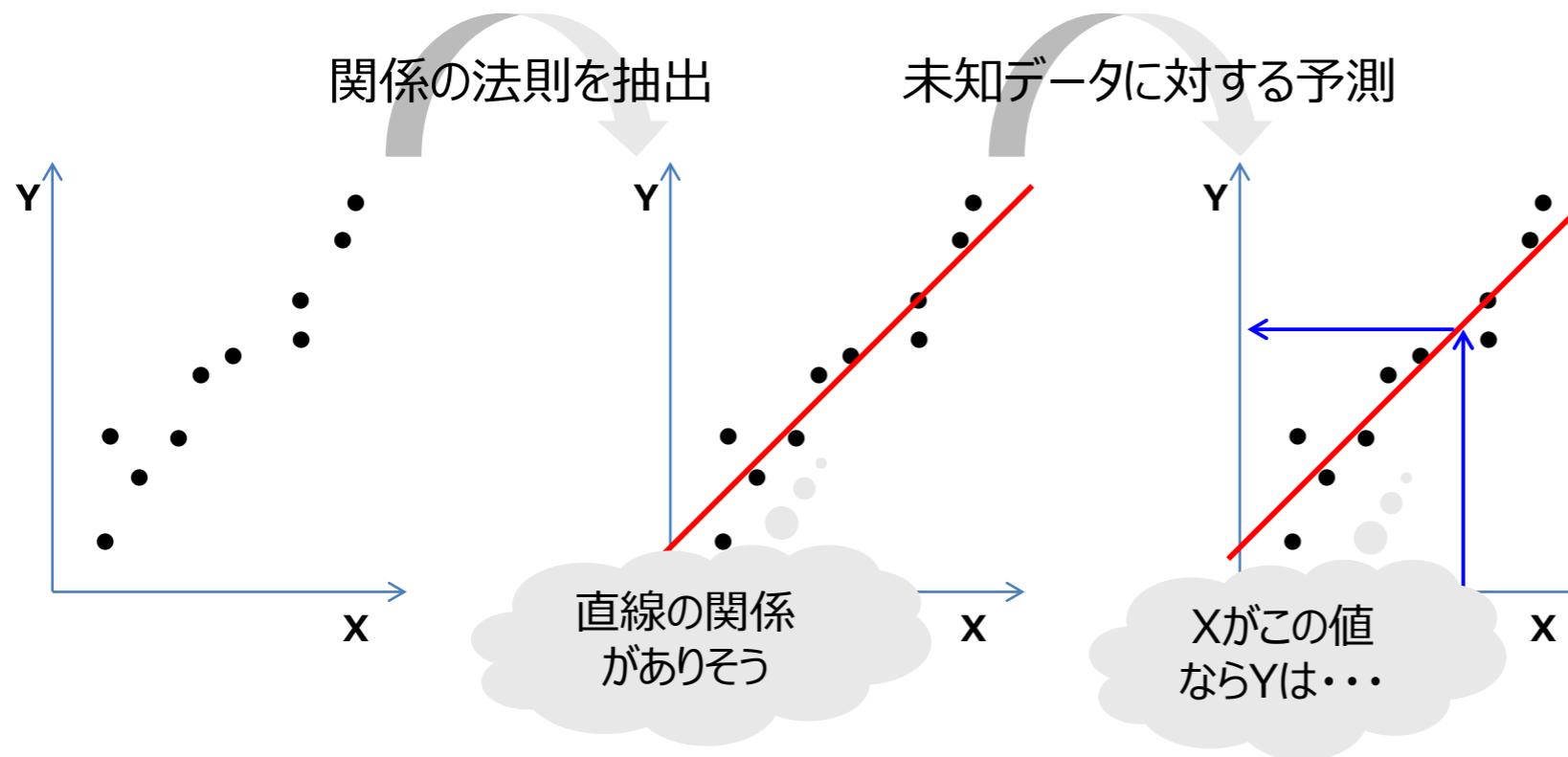
➤ 需要予測なしに、経営資源の再配分を含めた、仕入・生産・投資・販売などに関する有用なアクションプランを立案することができるだろうか？



予測を立ててから、アクションを起こそう

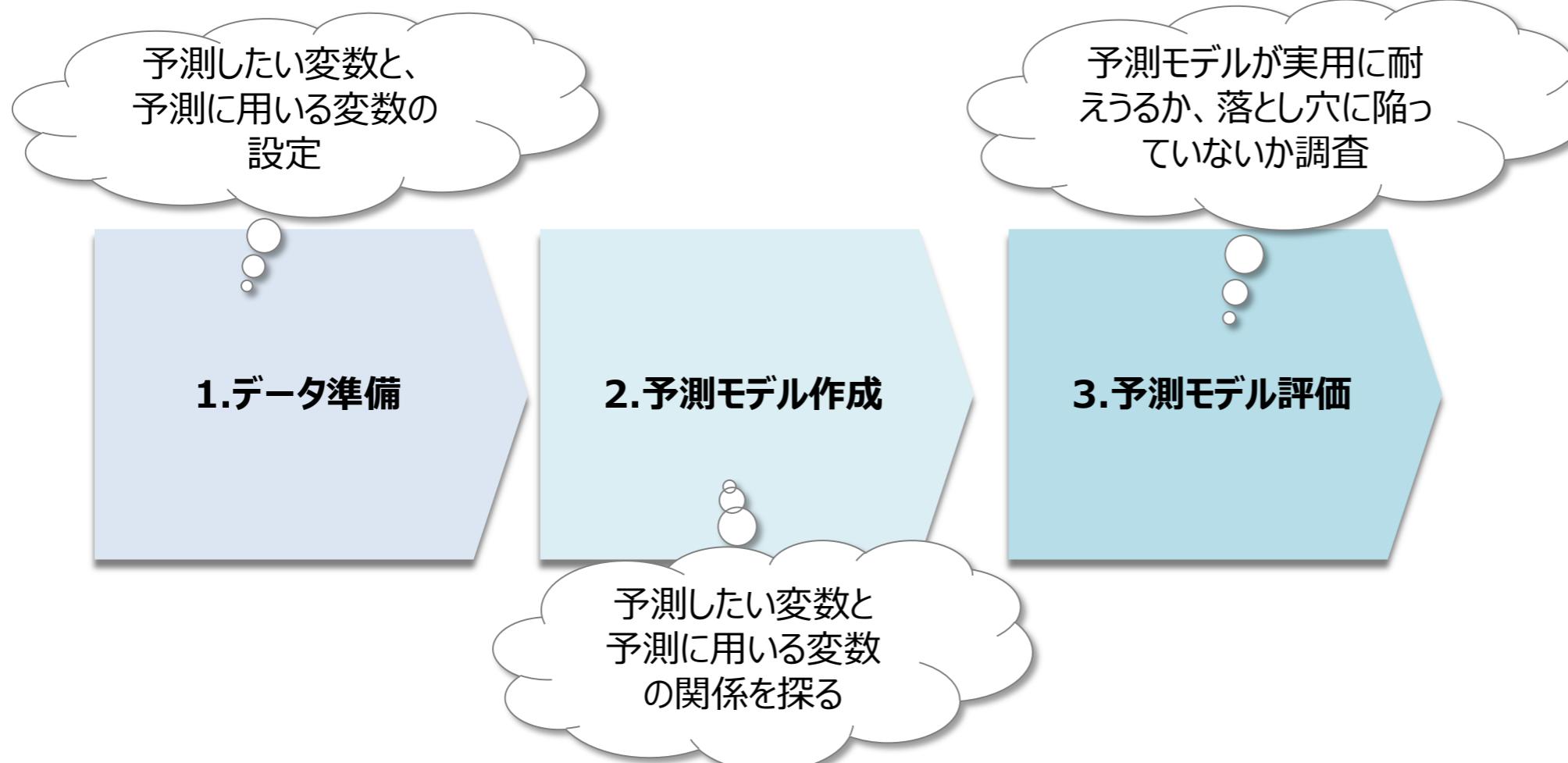
予測モデルとは

- 過去のデータから抽出された変数間の関係の法則
- 未知のデータに対し、抽出した法則を用いて予測を行うこと
ができる



予測モデル作成のプロセス

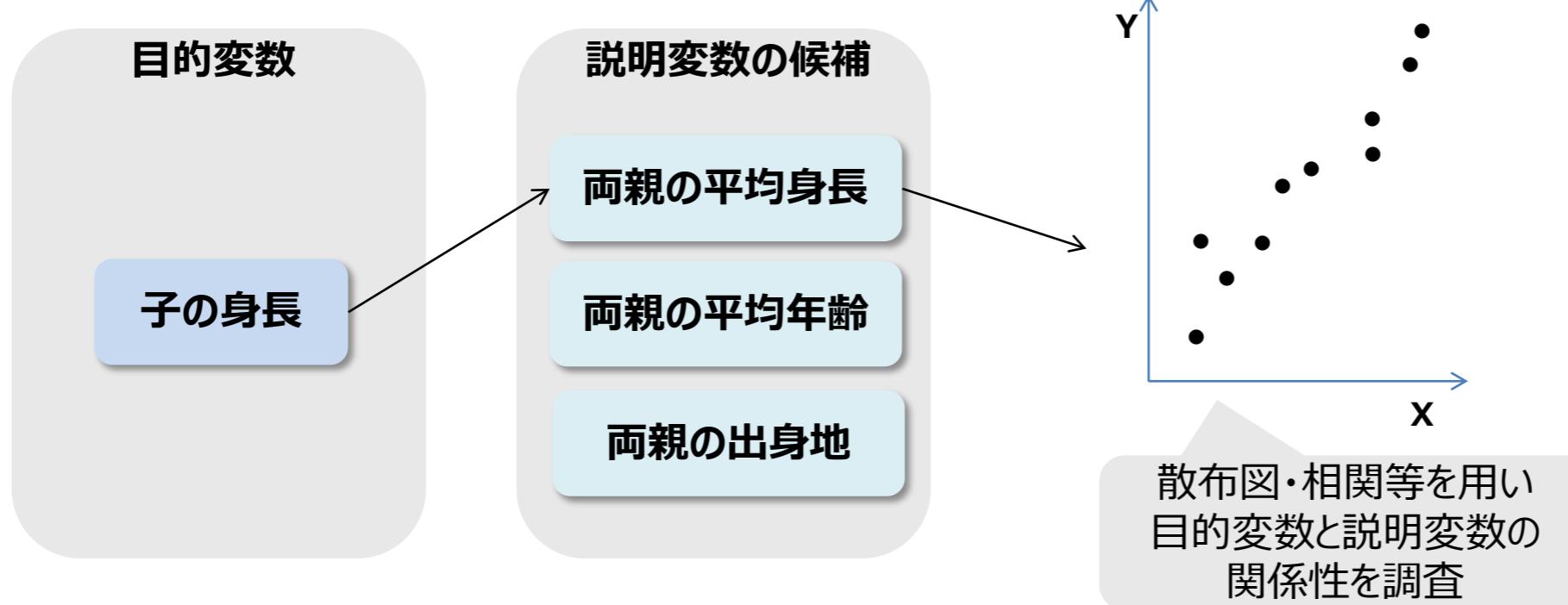
- 予測モデル作成は下記プロセスを経る必要がある



1.データ整備：目的変数・説明変数

- 予測したい変数を**目的変数**、
予測に用いる変数を**説明変数**と呼ぶ

- 説明変数は、**目的変数と関係があるものを選ぶ**必要がある
- 子の身長を予測したい場合…

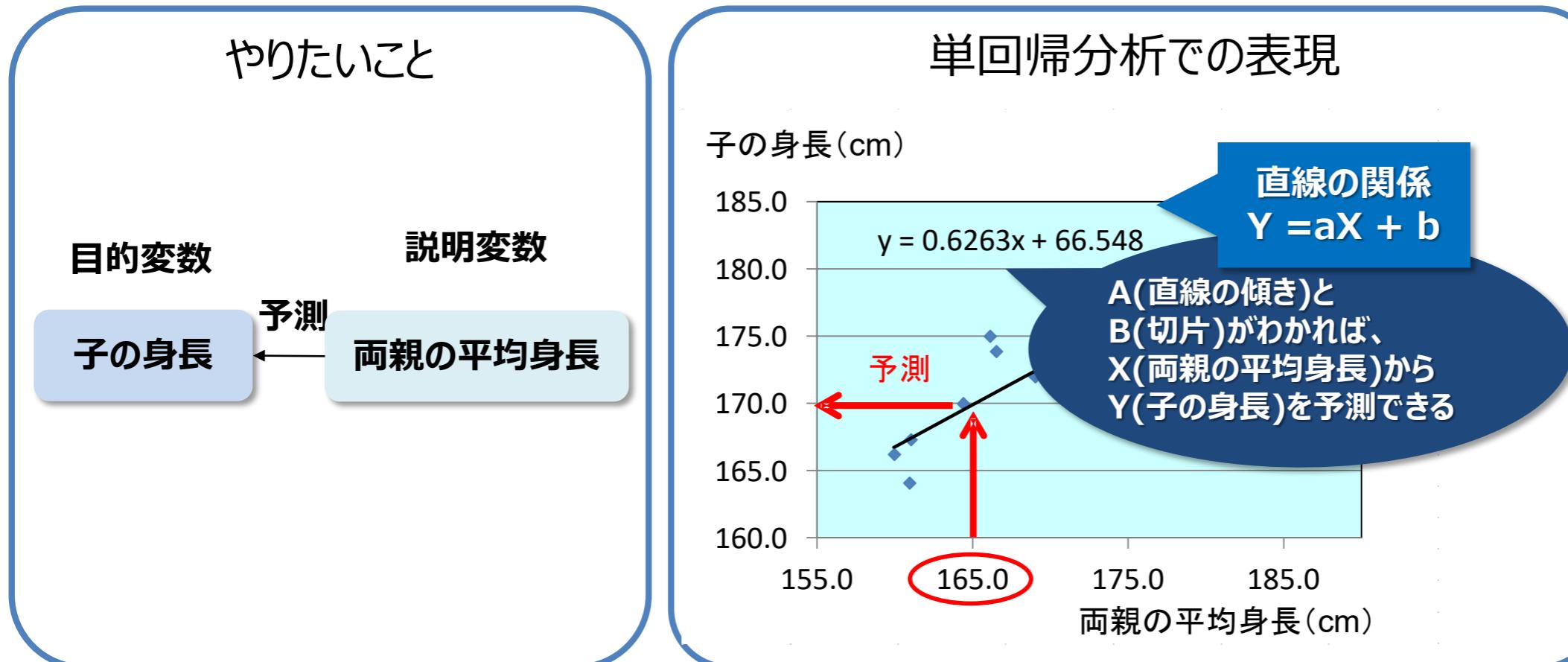


説明変数は目的変数と関係があるものを選ぶことが重要

2. 予測モデル作成：単回帰分析

- 目的変数と説明変数を直線の関係で表す

- 両親の平均身長から、子の身長を予測したい場合…



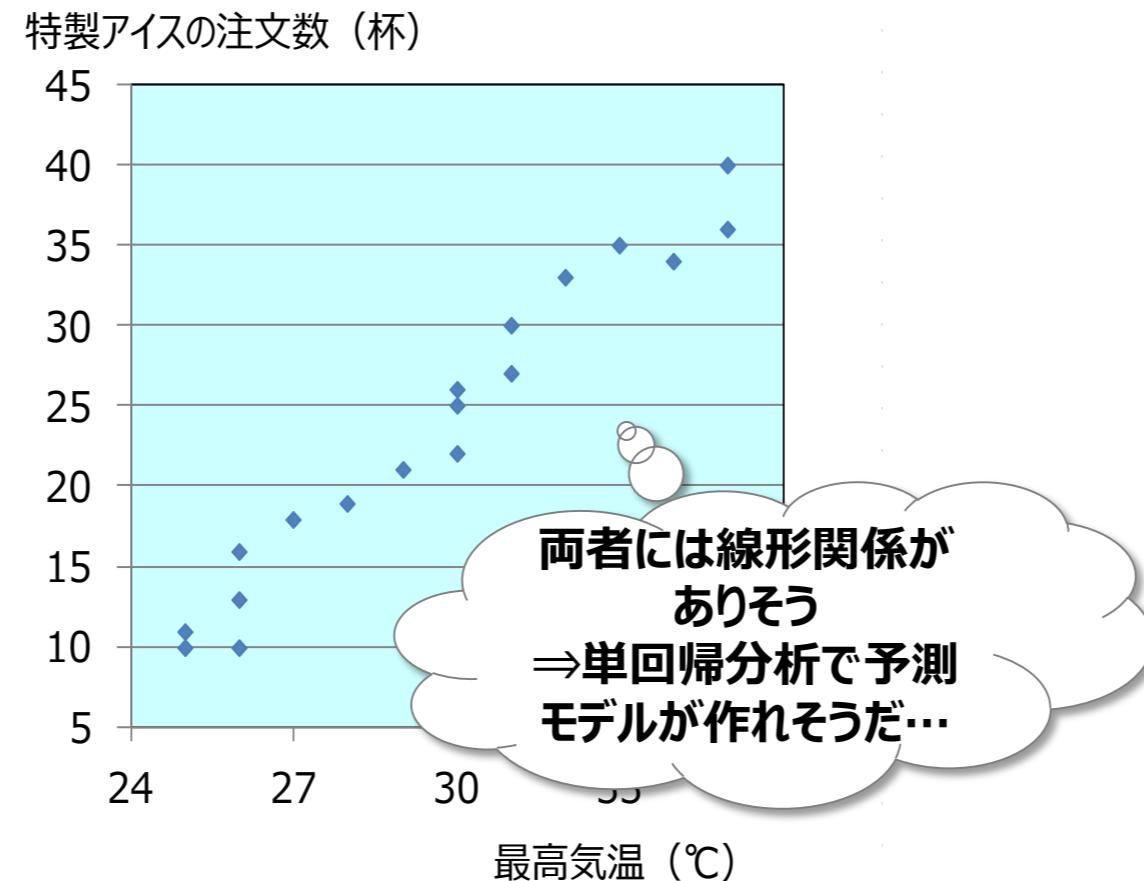
問題

- 右表は、最高気温とアイス注文数
- 仕込み量把握のため、翌日の注文数を予測したい
- 翌日の予想最高気温は30°C。アイスは何個準備しておくべきか？

最高気温 (°C)	特製アイスの注文数 (杯)
25	10
25	11
26	13
26	10
26	16
27	18
28	19
29	21
30	25
30	26
30	22
31	27
31	30
32	33
33	35
34	34
35	36
35	40

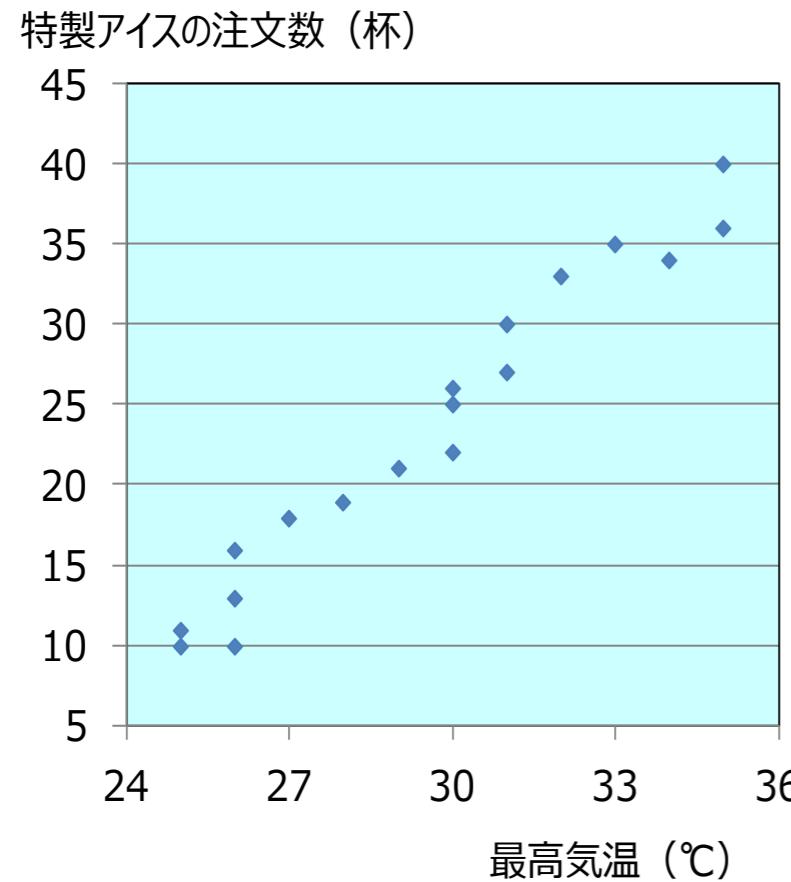
回帰分析の例：目的変数と説明変数の関係

- ①データ準備：
目的変数(アイス注文数)と説明変数(最高気温)の
散布図を描画し、線形関係があるか調査

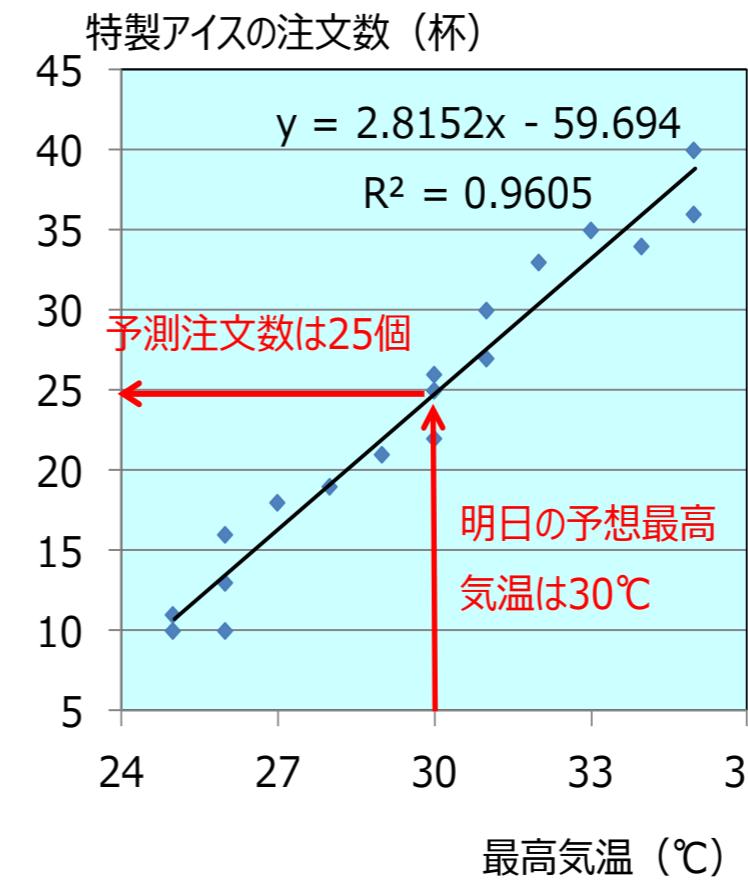


回帰分析の例：単回帰分析

- ②単回帰分析を実施(※具体的な方法は次回説明)
明日の予想最高気温を元に、注文数を予測



単回帰分析で
予測モデルを
算出



次回のテーマ

次回は

「モデル評価と予実評価」

お疲れ様でした！

社会人ためのデータサイエンス演習

第4週:ビジネスにおける予測と分析結果の報告

第2回:モデル評価と予実評価

講師名 矢島 安敏

第4週の内容紹介

第1回

- 回帰分析による予測

第2回

- モデル評価と予実評価

第3回

- 分析結果の報告 (記述/可視化方法)

第4回

- 分析結果の報告 (解釈の注意点)

第5回

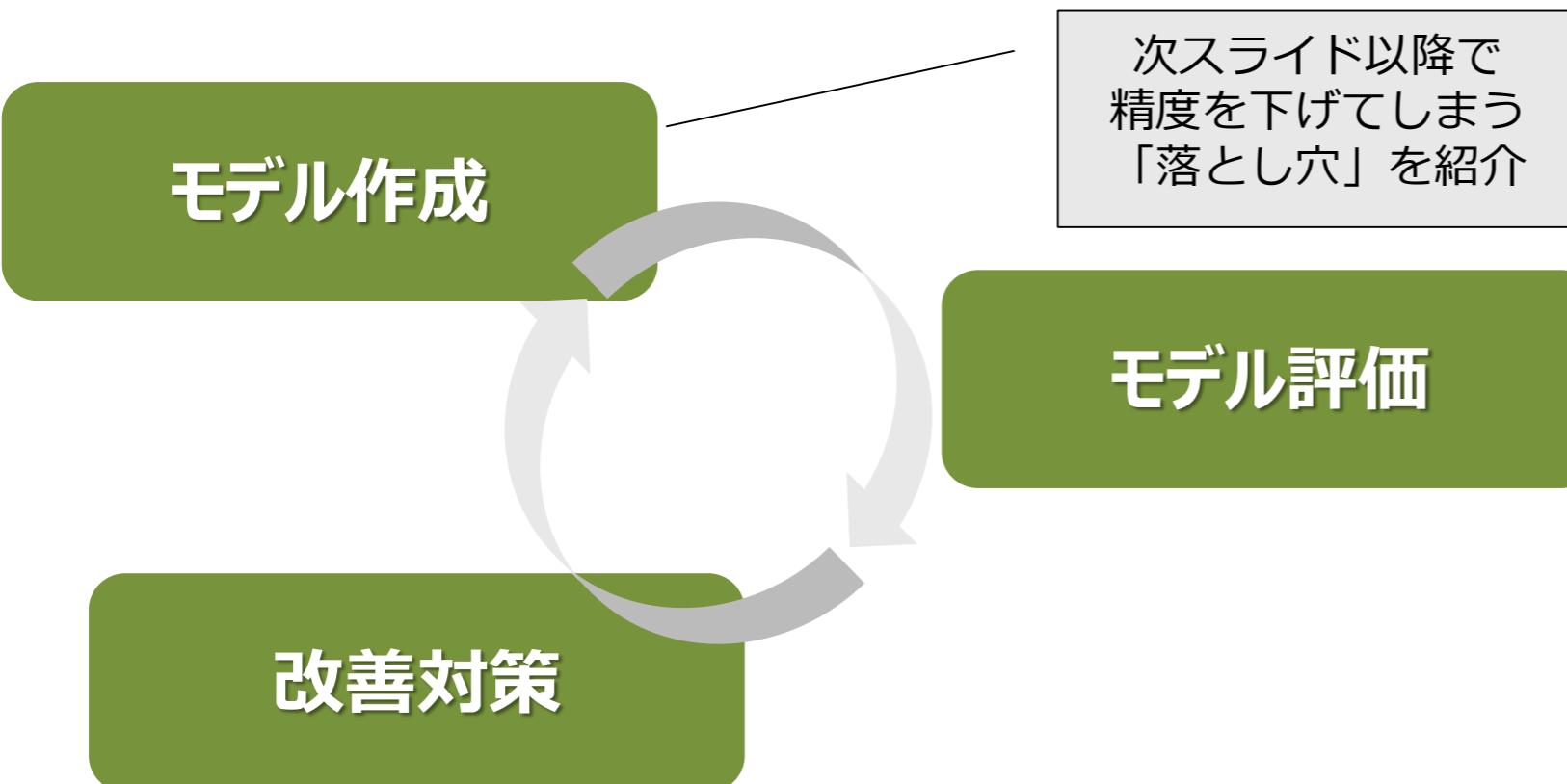
- 予測・分類等代表的手法と活用場面

第6回

- ビジネスシーンにおける「統計的検定」とその活用例

モデルの評価は何故必要なのか？

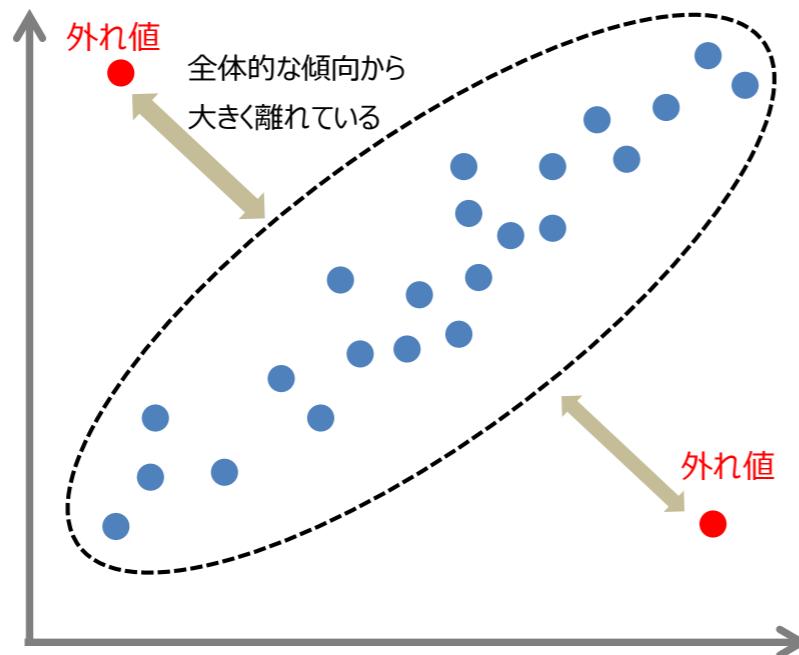
- 作成したモデルが信用できるか判断するために評価を実施
- 実務上は、モデル作成⇒評価⇒改善というフローを行う



誤った意思決定を避けるため評価を実施

落とし穴① 外れ値

- 大多数のデータから大きく離れた値は外れ値と呼ばれ、予測モデルの精度に悪影響を及ぼす
 - 外れ値があった場合は、**外れ値を削除して予測モデルを作成する必要がある。**



外れ値は除外して予測モデルを作成

落とし穴② 欠損値

- データに欠損値があると、予測モデルの精度に悪影響を及ぼす

- 欠損値のある場合は、
 - ・欠損値のあるデータ(レコードあるいは変数)は除外する
 - ・欠損値を適切な値で補完する いずれかの事前処理をすることが望ましい。

	身長 (cm)	体重 (kg)	年齢
Aさん	183.0	78	25
Bさん	167.4	52	49
Cさん	175.0	65	49
Dさん	152.4	48	欠損値
Eさん	173.9	71	59
Fさん	170.0	60	18
Gさん	179.3	85	36
Hさん	172.0	60	24

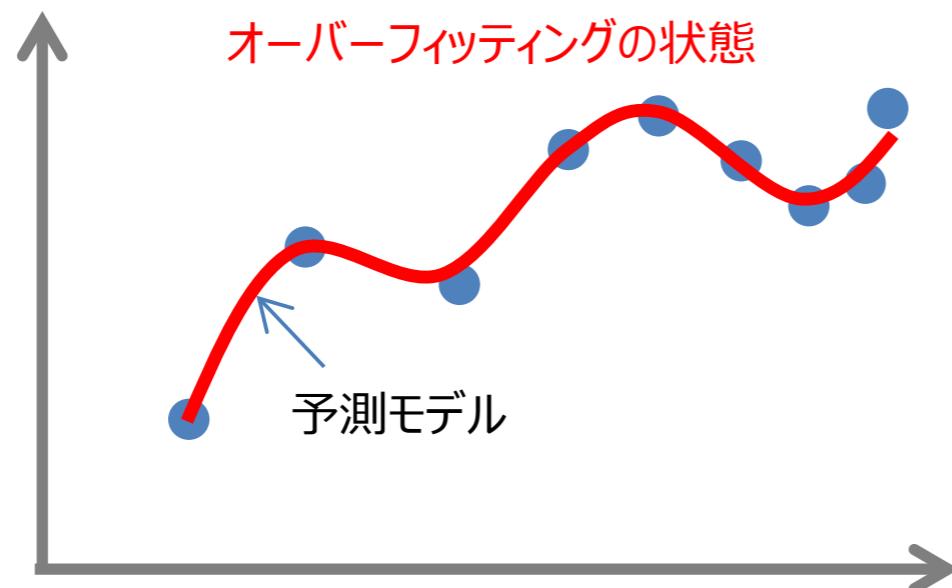
年齢は補完できない
ので、Bさんのデータ
は除外した方が
よさそうだ…

Fさんの体重は補完し
た値を使用しよう…

欠損値は除外、もしくは、補完して予測モデルを作成

落とし穴③ オーバーフィッティング

- オーバーフィッティング：当てはまりは極めて高いが…
 - 予測モデル作成に用いたデータに過剰にフィットしてしまい、未知のデータ(将来)に対する予測精度が悪くなってしまうこと

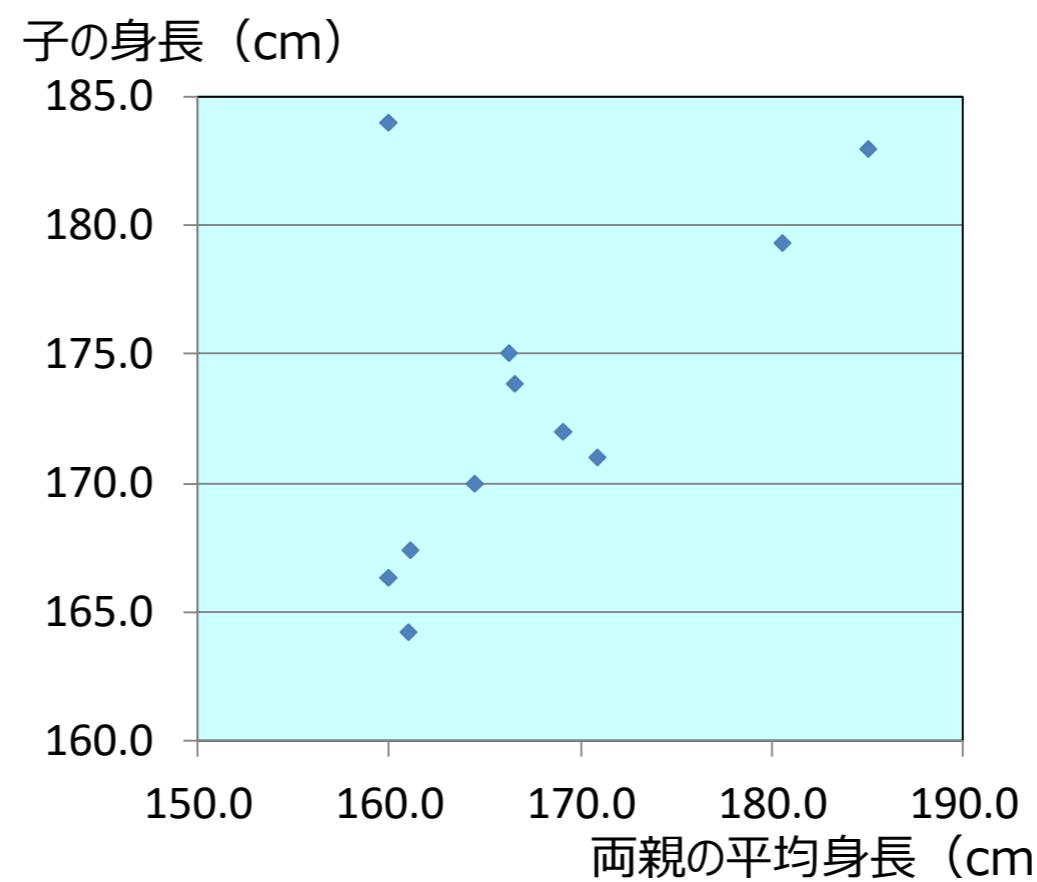


あてはまりの良すぎる予測モデルは
オーバーフィッティングを疑う

問題

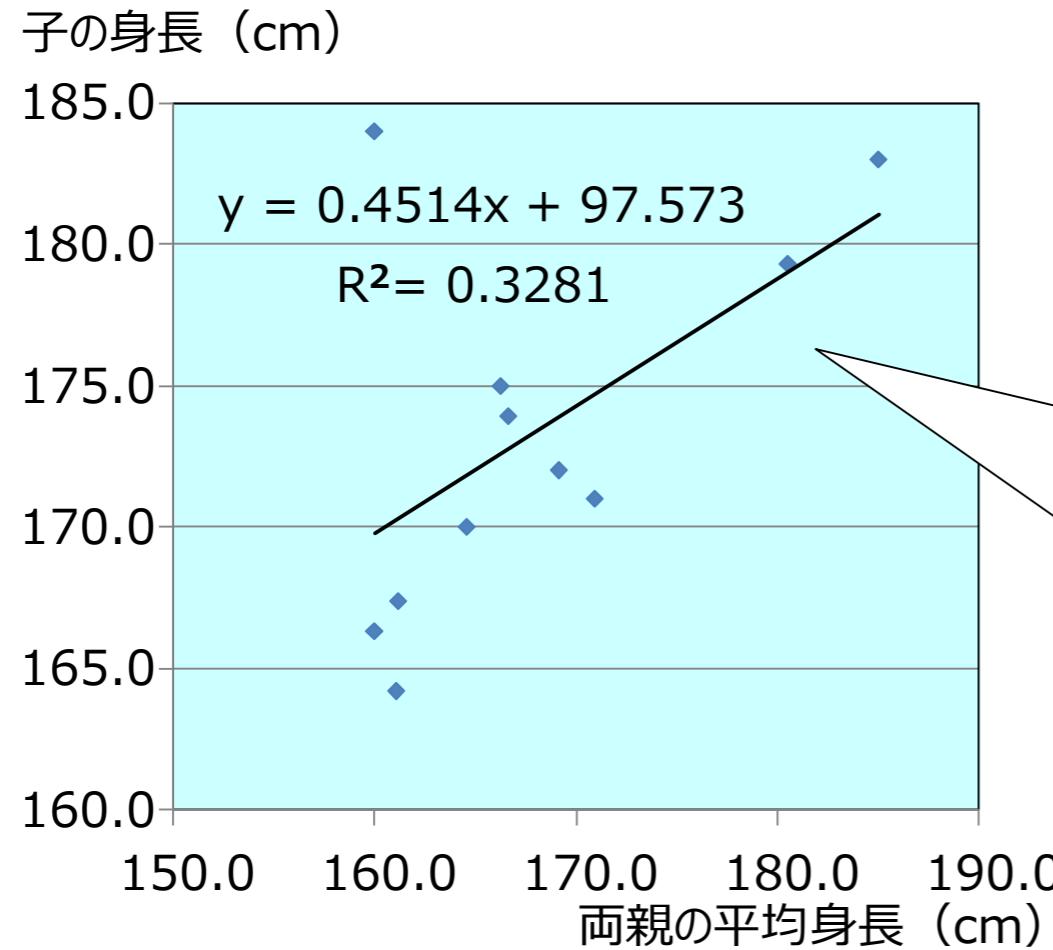
- 両親の平均身長から子の身長を予測するモデルを作成
※何に留意し、予測モデルを作成すべきか？

両親の平均身長	子の身長	(cm)
185.0	183.0	
161.1	167.4	
166.2	175.0	
161.0	164.2	
166.6	173.9	
164.5	170.0	
180.5	179.3	
169.1	172.0	
170.9	171.0	
160.0	166.3	
160.0	184.0	



回答①

- データをそのまま使用して予測モデルを作成

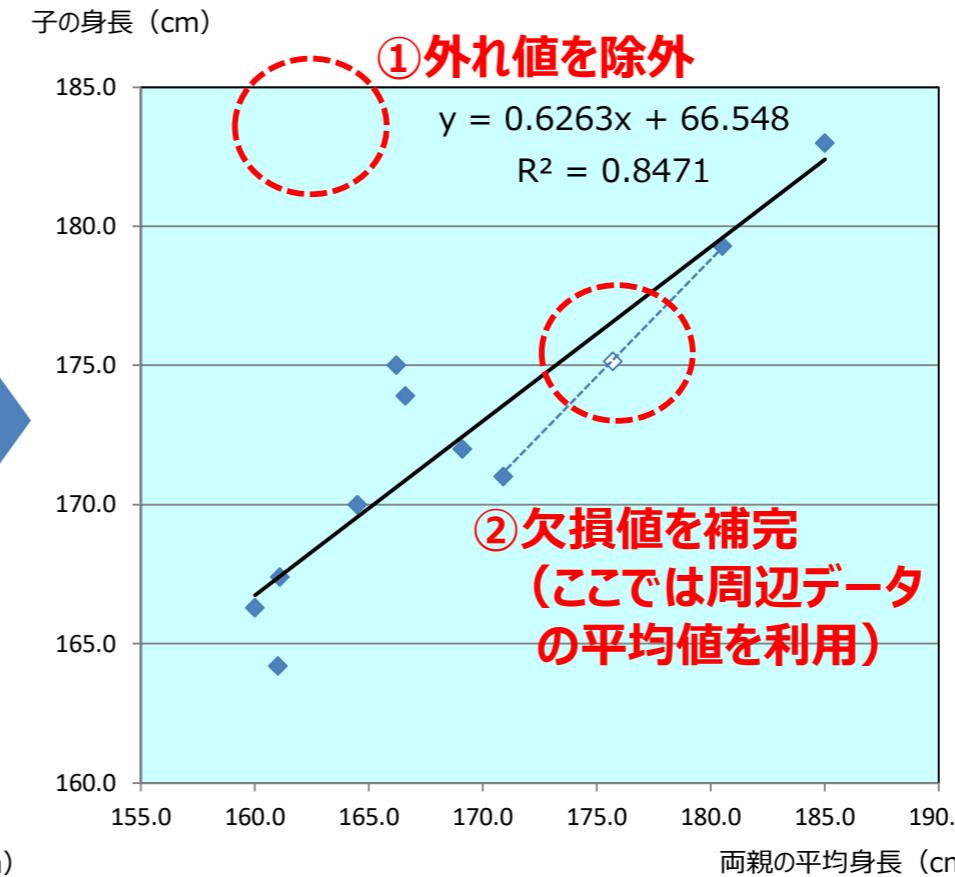
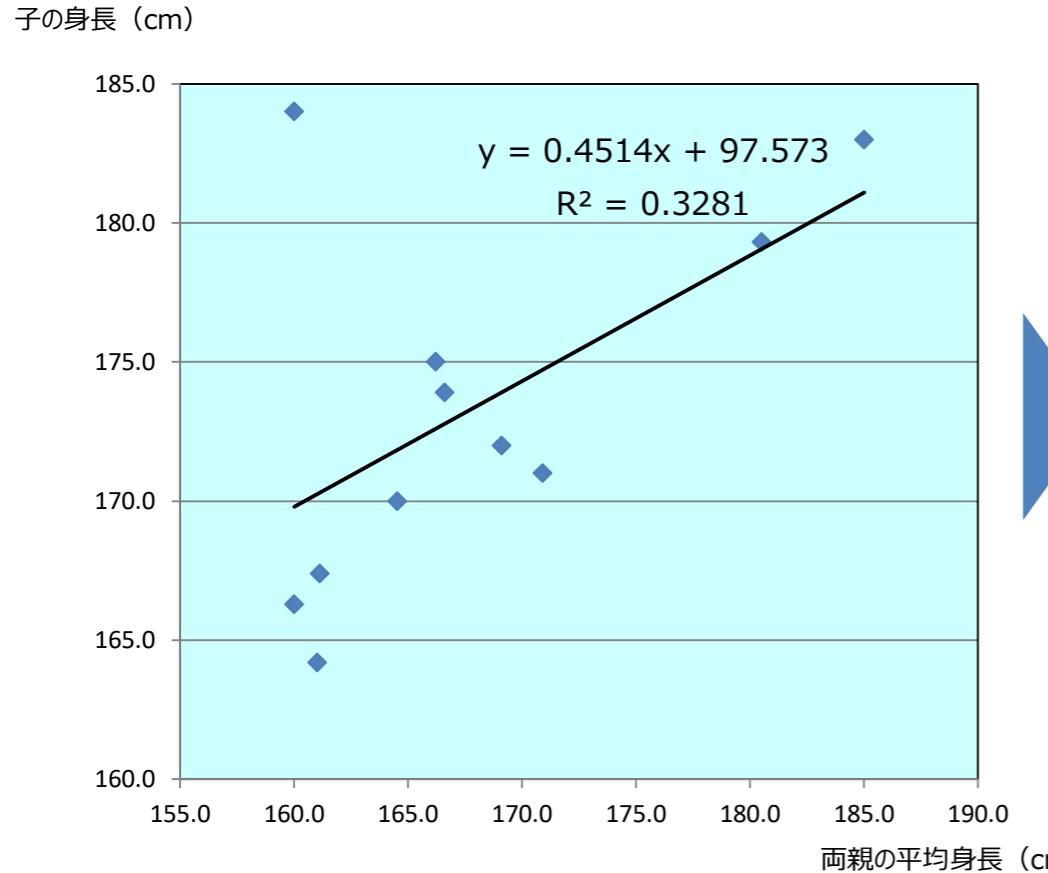


R²は、モデルの当てはまりを表す評価指標。現実のデータに対し、モデルが説明できる割合を表す(0 ~ 1の値を取り、高いほど良い)

予測モデルの当てはまり指標(R²)は0.32程度

回答②

- 外れ値を除外し、欠損値を補完したデータで予測モデルを作成



➤ モデルの決定係数(R^2)は0.32から0.85に上昇していることがわかる

予測モデルの当てはまり指標(R^2)が大幅に向

次回のテーマ

次回は

「分析結果の報告(記述/可視化方法)」

お疲れ様でした！

社会人のためのデータサイエンス演習

第4週:ビジネスにおける予測と分析結果の報告

第3回:分析結果の報告(記述/可視化方法)

講師名:高橋 範光

第4週の内容紹介

第1回

- 回帰分析による予測

第2回

- モデル評価と予実評価

第3回

- **分析結果の報告（記述/可視化方法）**

第4回

- 分析結果の報告（解釈の注意点）

第5回

- 予測・分類等代表的手法と活用場面

第6回

- ビジネスシーンにおける「統計的検定」とその活用例

報告の重要性

- 報告が不適切だと、正しい分析を行い有用な結果が得られてもミスリードを招く

「60歳代の女性」をターゲットにした新商品について、60代女性にアンケート調査をしたところ、店頭ではなくネットで買いたい人が8割を占めた。

ネットを主力と判断し、ネットと店頭の販売数を、8:2に設定した。

ネットの商品はほとんど売れなかつた！

実は、
調査方法が
Web調査

重要なのは、
得られた情報を正確に伝えるスキル

報告の要件

● 要件①調査分析の前提条件の明示

- 目的、用語の定義
- データの取得方法・出所(期間、対象者、データソース等)
- その他外在的要素、状況(社会情勢、制約条件等) など

● 要件②プロセスの明示

- 分析ロジック
- 調査分析フロー など

● 要件③適切な表現

- 情報を正確に伝える指標設定
- 示したい事柄に適した表、グラフの種類
- 図表の部位の明記(タイトル、軸、単位、凡例、出典等)

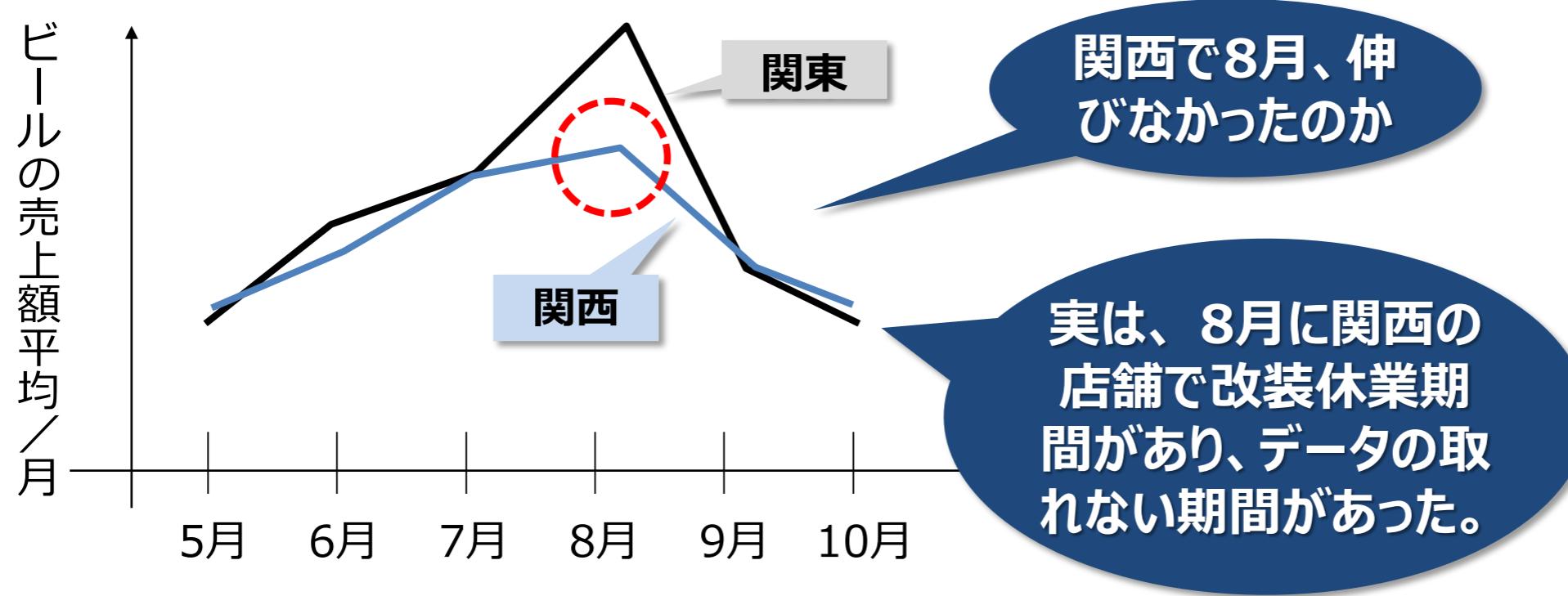
問題①

- 各店舗の毎月のビールの売上高を集計し、月別の売上額平均値の推移を上司に報告したいと思います。
- 上司は、関東と関西の売上の違いを見たがっていたので、それを中心に報告する予定です。
- どのような報告をすれば良いのでしょうか。

要件①調査分析の前提条件の明示

その他外在的要素、状況(社会情勢、制約条件等) など

- 関東と関西の売上額平均の推移を報告



→「店舗の休業期間」の報告漏れ

知らないと、結果の解釈が変わってしまうような
内容を明示することが必要

問題②

- 「60代女性」をターゲットとした新商品について、60代女性をターゲットとしたアンケート調査を行いました。
- 調査の目的は、新商品の許容価格帯を把握すること。
- どのような報告をすれば良いのでしょうか。

要件②プロセスの明示

調査分析フロー など

- アンケート調査では調査手法によって、回答結果が異なる可能性

~~インターネット
調査~~



→インターネット調査での「60代女性」の回答結果は代表性に注意が必要

データ収集方法や分析手法などのプロセスが分析結果に影響を及ぼす場合は、事前に提示

要件③適切な表現

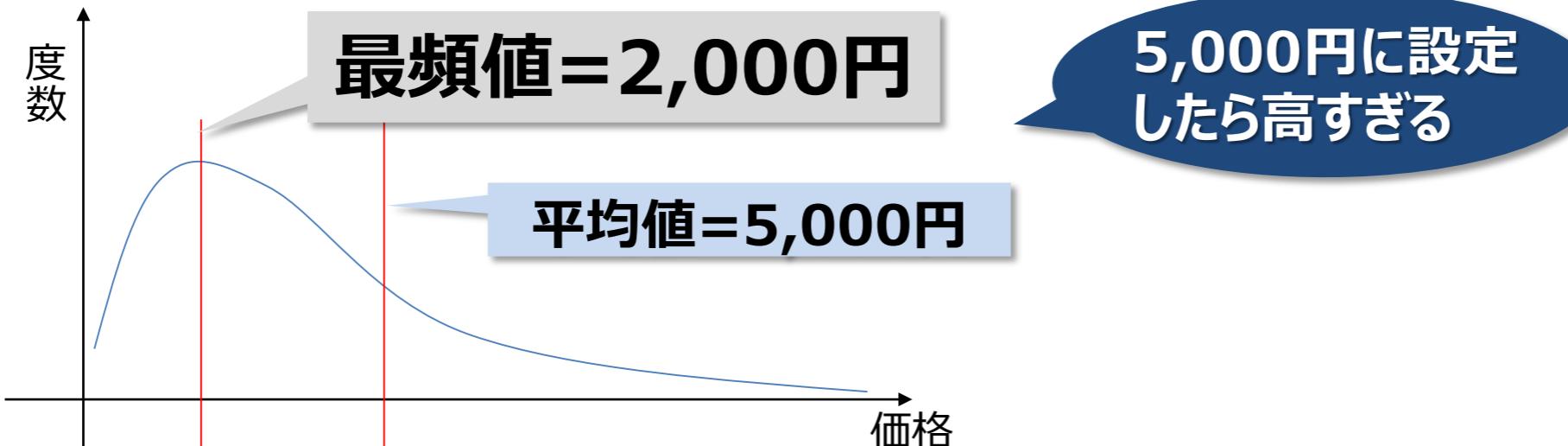
情報を正確に伝える指標設定

● 調査結果における新商品の許容価格を報告

回答における許容価格の
平均値 = 5,000円

新商品は、
5,000円前後に

ところが、グラフを描いてみると、

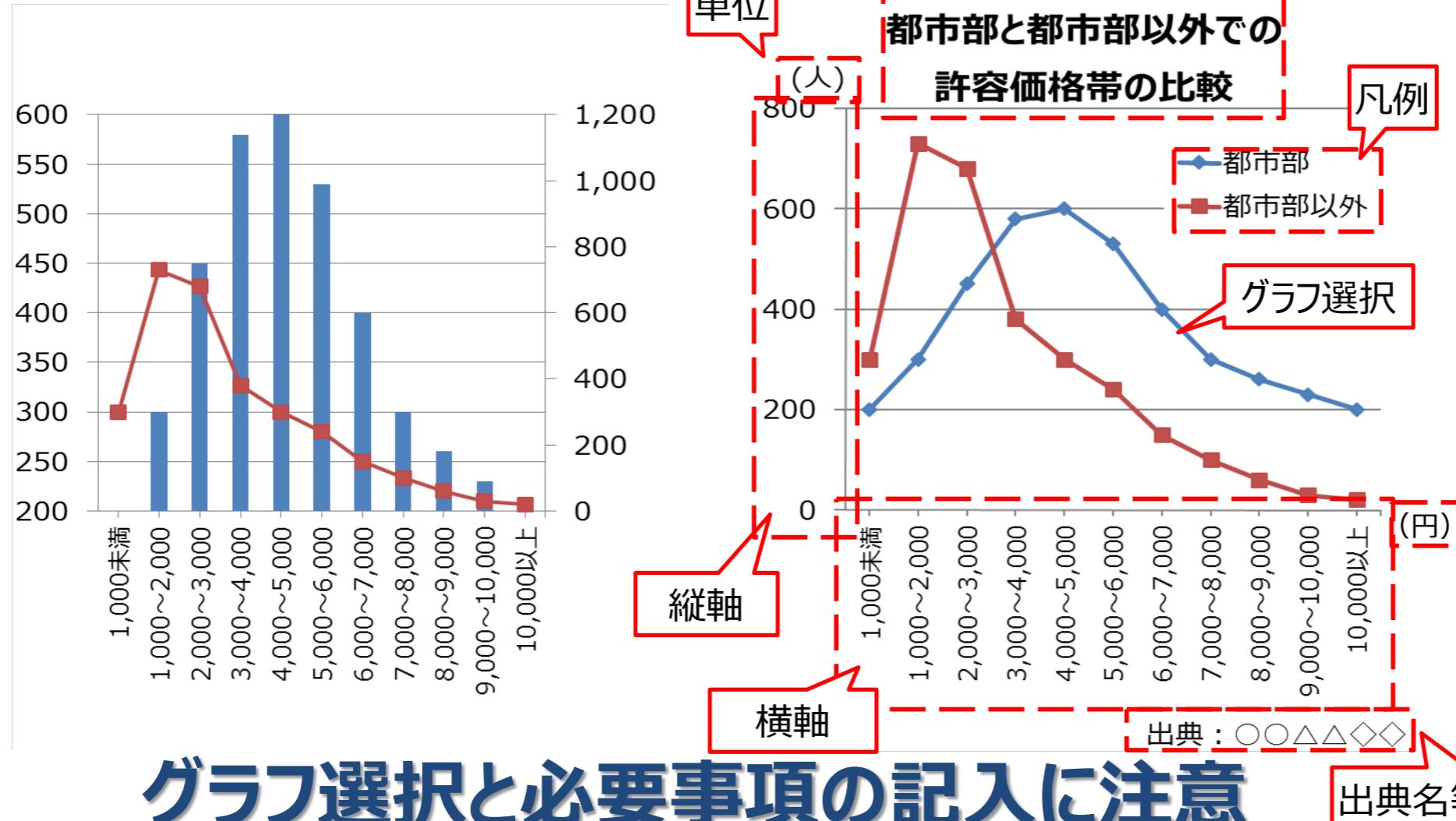


求められる分析課題に対して、
適切な指標を設定し提示することが重要

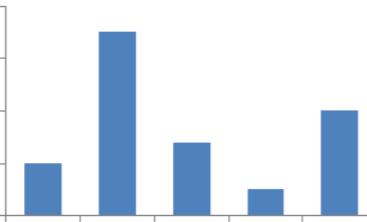
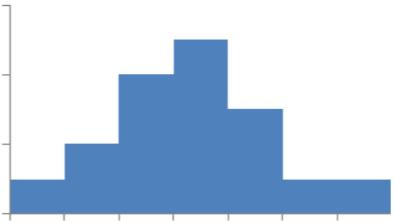
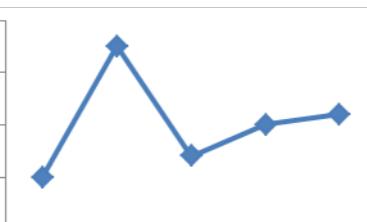
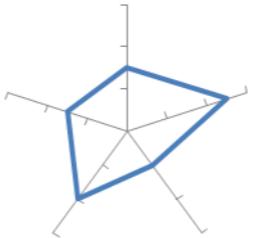
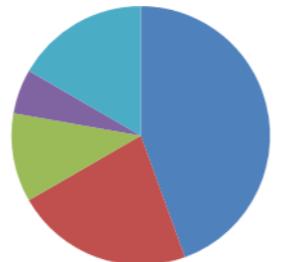
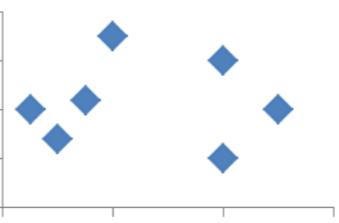
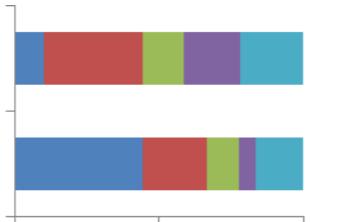
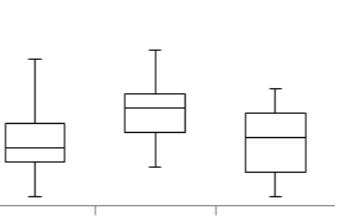
要件③適切な表現

示したい事柄に適した表、グラフの種類 / 図表の部位の明記

- 都市部と都市部以外で許容価格帯に差異があるか報告



グラフの種類と用途

棒グラフ	棒の長さで、量の大きさを比較する。		ヒストグラム	柱の面積で、データの散らばり具合を見る。	
折れ線グラフ	線の傾きで、ある数量の連続的な変化を見る。		レーダーチャート	複数の指標をまとめてみる。	
円グラフ	パイの面積で、全体に占める構成比を見る。		散布図	2種類のデータの相関を見る。	
帯グラフ	全体に占める構成比を比較する。		箱ひげ図	平均値等の指標でデータの散らばり具合を見る。	

次回のテーマ

次回は

「分析結果の報告(解釈の注意点)」

お疲れ様でした！