情報リテラシー(第10回 後期)ハンドアウト

データ分析と仮説検証サイクル

1. 今日のねらい

- **仮説検証サイクル**の考え方を理解する
- 散布図と**相関**の関係を理解する
- Pythonで**線形回帰**を実装できる
- 予測モデルの**可能性と限界**を知る

---

2. 仮説検証サイクルとPDCA

仮説検証サイクル(科学): 仮説 → 検証 → 分析 → 結論 → 新仮説

PDCA(ビジネス): Plan → Do → Check → Act

共通点: 一度で完璧を目指さず、繰り返し改善する

4ステップ: ①仮説(「〇〇なら△△になるはず」と予想)、②検証(データで確かめる)、③分析(結果を詳しく調べる)、④結論(仮説は正しい？修正が必要？)

---

3. 今日の仮説

仮説: 「打率が高い選手ほど、本塁打も多い」

検証方法: 散布図で視覚的に確認／相関係数で数値的に確認／線形回帰で予測モデルを作成

データ: 野球選手40人分の成績データ(架空)。セ・リーグ20人、パ・リーグ20人。項目：選手名、リーグ、打率、本塁打、打点、安打、出塁率

---

4. 散布図と相関係数

散布図: 2つのデータの関係を見るグラフ

相関係数: 相関の強さを数値化(-1 〜 +1)。今日のデータでは約0.7前後 → 比較的高い正の相関

注意点: 相関 = 因果ではない。相関が高くても「打率が本塁打を増やしている」とは断定できない。他の要因(体格、スイング、戦術など)が関係する可能性がある

---

5. 線形回帰と予測

線形回帰: 散布図に「最も当てはまる直線」を引く方法

数式: $y = a x + b$ (x:説明変数(打率)、y:目的変数(本塁打)、a:傾き、b:切片)

決定係数(R^2): 今回のデータでは約0.5〜0.6 → モデルが説明できるのは約半分程度

予測例: 打率.300の選手 → 約15〜20本程度／打率.250の選手 → 約10〜15本程度

外れ値: 打率低いのに本塁打多い選手、打率高いのに本塁打少ない選手 → 例外の発見につながる

---

6. 訓練データとテストデータ

なぜ分ける？: モデルが未知のデータでも予測できるか確認するため

訓練データ: モデルの学習に使う(60%)。このデータでパラメータを決める

テストデータ: モデルの評価に使う(40%)。未知のデータとして扱う

重要: 訓練データで学習 → テストデータで評価 → 汎化性能を確認

---

7. Colabの開き方

手順

1. Teamsの課題からリンクをクリック
2. 「baseball_data.csv」を自分のマイドライブにコピー(右クリック → 「ドライブにコピーを作成」)
3. Colabノートブックを開く(「ドライブにコピーを保存」をクリック)
4. セル1を実行(Googleドライブへのアクセス許可が必要)
5. セル2〜6を順番に実行

実行順序: ①データ読み込み → ②散布図 → ③相関係数 → ④train/test分割 → ⑤線形回帰 → ⑥予測と可視化

---

8. 提出課題

課題: データ分析実習

提出物: Wordファイル1つ

内容: ①セル6の実行結果のスクリーンショットを貼る(散布図、予測結果、外れ値トップ3)／②外れ値が出た理由を1行で書く(例:「打撃スタイルの違いによるもの」)

提出方法: WordファイルをTeamsに提出

---

9. まとめ(キーワード)

仮説検証サイクル／PDCA／散布図／相関係数／相関と因果／線形回帰／説明変数／目的変数／決定係数(R^2)／外れ値／訓練データ／テストデータ／汎化性能