第6章 重回帰分析

(Multiple Regression Analysis)

HDS 中村 知繁

T. Nakamura | Juntendo Univ.

2025/02/08

学習目標

この章を学ぶことで、以下の能力を習得します。

- 複数の説明変数を用いて目的変数を予測する重回帰モデルを定式化し、解釈する。
- 数値変数とカテゴリ変数が混在するデータに対して、交互作用モデルと平行傾斜モデルを構築し、その違いを理解する。
- 回帰平面の概念を理解し、2つの数値説明変数を持つモデルを可視化・解釈する。
- モデル選択の基本的な考え方を理解し、可視化や決定係数 R^2 を用いてモデルの適合度を評価する。
- シンプソンのパラドックスのような、多変量データに潜む統計的現象を特定し、説明する。

導入

第5章では、単一の説明変数で結果をモデル化する単純線形回帰を扱いました。
しかし、現実世界の現象は、単一の要因だけでは説明できません。

重回帰分析 (Multiple Regression Analysis) は、複数の説明変数 $x_1, x_2, \ldots, x_p$ を同時に考慮に入れて目的変数 y をモデル化する強力な手法です。

複数の変数を同時に分析することで、各変数が他に与える影響を調整した上での「純粋な」関連性を評価できます。これは、交絡（confounding）による見せかけの相関を排し、より本質的な関係性を探る上で不可欠です。

必要なライブラリ

本章の分析では、以下のPythonライブラリを使用します。

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LinearRegression
# 決定係数や平均二乗誤差の計算に使用
from sklearn.metrics import r2_score, mean_squared_error
# 3Dプロット用
from mpl_toolkits.mplot3d import Axes3D
```

6.1 1つの数値変数と1つのカテゴリ変数を持つモデル

`seaborn` の `tips` データセットを用いて、チップ額をモデル化します。

1. **目的変数 y (数値)**: チップ額 (`tip`)
2. **説明変数 (2つ)**:
    1. x_1 (数値): 会計総額 (`total_bill`)
    2. x_2 (カテゴリ): 支払い客の性別 (`sex`)

問い: 会計総額はチップ額にどう影響するか？また、その影響は性別によって異なるか？

6.2 2つの数値変数を持つモデル

`seaborn` の `penguins` データセットを用います。

1. **目的変数** y: ペンギンの体重 (`body_mass_g`)
2. **説明変数**:
    1. x_1: くちばしの長さ (`bill_length_mm`)
    2. x_2: フリッパーの長さ (`flipper_length_mm`)

6.2.1 探索的データ分析

1. データ準備と相関の確認

```
penguins = sns.load_dataset("penguins")
penguins_ch6 = penguins[['body_mass_g', 'bill_length_mm', 'flipper_length_mm']].dropna()

# 相関行列を計算
correlation_matrix = penguins_ch6.corr()
print(correlation_matrix)

# 相関行列をヒートマップで可視化
plt.figure(figsize=(7, 5))
sns.heatmap(correlation_matrix, annot=True, cmap='viridis', vmin=0, vmax=1, fmt=".3f")
plt.title('変数間の相関ヒートマップ')
plt.show()
```

観察:
体重はフリッパー長と強い正の相関 ($r = 0.873$)、くちばし長とも中程度の正の相関 ($r = 0.595$) があります。

6.2.2 回帰平面 (Regression Plane)

2つの数値説明変数を持つモデルは、3次元空間の**平面**として表現されます。

```
X = penguins_ch6[['bill_length_mm', 'flipper_length_mm']]
y = penguins_ch6['body_mass_g']

penguin_model = LinearRegression().fit(X, y)
intercept_p = penguin_model.intercept_
coefs_p = penguin_model.coef_

print(f"切片 (b_0): {intercept_p:.4f}")
print(f"bill_length_mm (b_bill): {coefs_p[0]:.4f}")
print(f"flipper_length_mm (b_flipper): {coefs_p[1]:.4f}")
```

回帰平面の方程式:

$$\widehat{\text{weight}} = -5980.6869 + 33.3499 \cdot \text{bill\_length} + 44.9101 \cdot \text{flipper\_length}$$

係数の解釈:

- `b_bill`: **フリッパー長を固定**した場合、くちばしが1mm長くなると体重は約33.3g増加。
- `b_flipper`: **くちばし長を固定**した場合、フリッパーが1mm長くなると体重は約44.9g増加。

6.3.1 モデル選択: 複雑さと適合度のトレードオフ

問い: 交互作用モデル vs 平行傾斜モデル、どちらを選ぶべきか？
→ **モデル選択 (Model Selection)** の問題

指導原理: オッカムの剃刀 (Occam's Razor)

> 「他の条件が同じなら、より単純なモデルが望ましい」

モデルの複雑さを増すのは、それがデータの構造を説明する上で**実質的な改善**をもたらす場合にのみ正当化されます。

`tips` データの例では、交互作用の効果は非常に小さかったため、より単純な**平行傾斜モデル**を選ぶのが合理的です。

6.3.2 決定係数 R^2 によるモデル評価

決定係数 R^2 (R-squared) は、目的変数 y の全変動のうち、モデルによって説明された変動の割合を示します。

$$R^2 = \frac{Var(\hat{y})}{Var(y)} = 1 - \frac{Var(e)}{Var(y)}$$

R^2 は 0 から 1 の値をとり、1に近いほど適合度が高いことを意味します。

`tips` データでの比較:

```
r2_interaction = r2_score(y, interaction_model.predict(X))
r2_parallel = r2_score(y_parallel, parallel_model.predict(X_parallel))

print(f"交互作用モデル R²: {r2_interaction:.4f}")   # -> 0.4574
print(f"平行傾斜モデル R²: {r2_parallel:.4f}") # -> 0.4574
```

R^2 の差はごくわずかであり、交互作用項を追加しても説明力はほとんど向上しません。これは、より単純な平行傾斜モデルを選択する判断を裏付けます。

6.3.3 シンプソンのパラドックスと交絡変数

シンプソンのパラドックス (Simpson's Paradox)

> データ全体で観測される傾向が、データをサブグループに分割すると消滅、あるいは逆転する現象。

`penguins` データでは、**種 (`species`)** が**交絡変数 (confounding variable)** となる可能性があります。

- **全体（黒破線）**: くちばしが長いほど体重が重い。
- **種ごと**: グループ内の傾きは全体より緩やか。

種という交絡変数を無視すると、くちばしの長さと体重の関係を過大評価してしまう可能性があります。

6.4 結論と次章への展望

本章のまとめ:

- **重回帰**: 複数の説明変数で目的変数をモデル化。
- **交互作用**: ある説明変数の効果が、別の変数の水準によって変化する現象。
- **モデル選択**: オッカムの剃刀を指針とし、複雑さと適合度（R^2）のバランスを考慮。
- **交絡**: 第三の変数が原因で、2変数間に見せかけの相関が生じる現象（シンプソンのパラドックス）。

次章への展望:
ここまでは、手元の**標本 (sample)** にフィットする線を「推定」してきました。
次章から始まるパートIII「**統計的推測**」では、標本から得られた知見を、より大きな**母集団 (population)** 全体に一般化するための理論と手法を学びます。