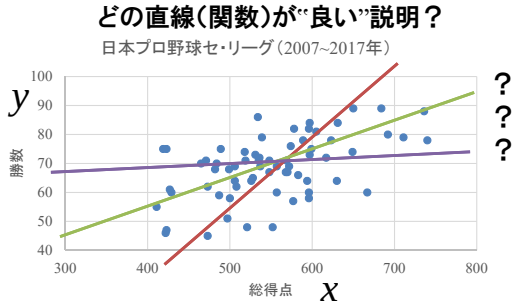
経済と社会 補助資料
～マーケティング・リサーチと回帰分析～

2022年度1学期：　火曜2限
担当教員：　石垣　司

回帰分析って？

- 変数 y の動きを，別の変数 x と関数 f で予測や説明
  - 最もポピュラーなデータ分析手法の一つ
- 線形回帰分析
  - 関数 f ：直線（1次式）
    $y = f(x) = a + bx$
    - 散布図のデータの散らばりに適合する直線（切片 a ,傾き b ）を推定
    - x と y に因果関係がある場合は，x が原因，y が結果の表現



- これ以降の講義の目的：
  線形重回帰分析の結果の意味を理解する

マーケティング・リサーチと回帰分析

- 需要の予測（小売業での例）
  - 曜日，天気，チラシ広告の有無などから来店人数を予測。過去の購買履歴データから算出されるPI値(Purchase Index:来店しレジを通過した顧客1000人当たりの各商品の購買指数)から商品の需要量を予測
- 市場反応分析（メーカーでの例）
  - 販売価格，プロモーションの有無などから売上数量を予測。交差価格弾力性などの経済学的指標を算出することで，ブランド間の競合関係を測定する

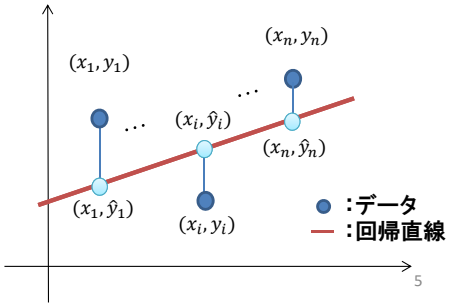
回帰分析の用語の整理

- 変数に関する用語
  - 説明変数　　変数 x
  - 従属変数　　変数 y（目的変数、被説明変数）
  - 回帰係数　　係数 b（パラメータ）
- 回帰に関する用語
  - 回帰　　ある変数を他の変数の関数で表現すること
  - 回帰式　　$y = a + bx$（単回帰式）
  - 回帰直線　　回帰式が表現する直線
  - 単回帰分析　　1つの説明変数を用いた回帰による分析
  - 重回帰分析　　複数の説明変数を用いた回帰による分析

回帰式の推定

- **データから回帰係数 b と切片 a を決定する**
  - **合理的な基準と手続きに基づいた推定が必要**
- **基準　残差平方和**(RSS: Residual sum of squares)**の最小化**
  - **残差**e_i　　$e_i = y_i - \hat{y}_i = y_i - (a + bx_i)$
  - **残差平方和**　　$RSS = \sum_{i=n}^{n} e_i^2$
- **手続き　最小2乗法**
- **最小2乗推定量**

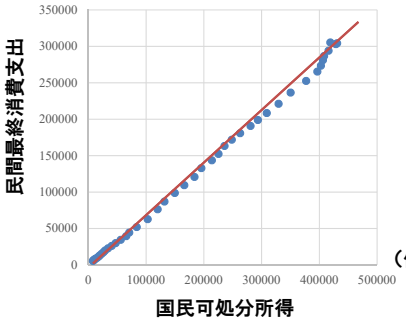
$$\hat{b} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2},\ \hat{a} = \bar{y} - \hat{b}\bar{x}$$



回帰分析と予測

- **回帰式 $y = \hat{a} + \hat{b}x$ を利用した予測**
  - $\hat{a}$ と $\hat{b}$ はデータ $\{(x_1, y_1), \cdots, (x_n, y_n)\}$ から推定された係数
  - $\hat{y}_{n+1}$：　x_{n+1}に対する 変数 y の予測値

$$\hat{y}_{n+1} = \hat{a} + \hat{b}x_{n+1}$$



1955年度～1998年度(1968SNA)
（内閣府　国民経済計算年次推計）

問題
左図の回帰係数は$\hat{a} \cong 0, \hat{b} \cong 0.7$である。国民可処分所得が500兆円のとき，民間最終消費支出の予測値は？

重回帰分析

- **複数の説明変数による回帰分析**
  - **目的変数，従属変数：　変数** y
  - **説明変数，独立変数：　変数** $x_1, x_2, \cdots, x_P$
  - **データ：**　　$\{y_i, x_{i1}, x_{i2}, \cdots, x_{iP}\}(i = 1, \cdots, N)$
  - **偏回帰係数：**　　**係数** $b_1, b_2, \cdots, b_P$（**パラメータ**）
  - **切片：**　　**係数** b_0（**パラメータ**）
- **重回帰式**

$$y = b_0 + b_1 x_1 +, \cdots, +b_P x_P$$

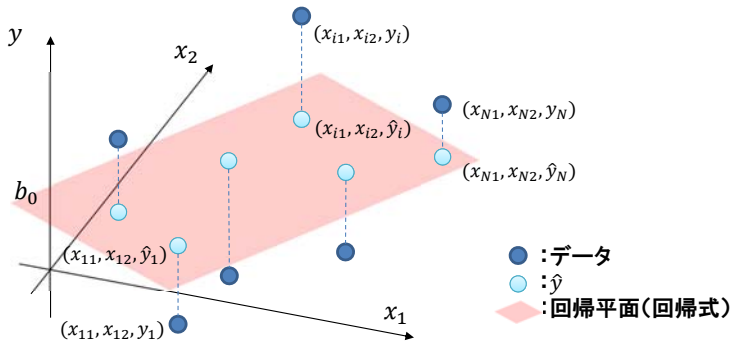
- **重回帰モデル**（**データによる記述**）

$$y_i = b_0 + b_1 x_{i1} +, \cdots, +b_P x_{iP} + e_i \ \ (i = 1, \cdots N)$$

重回帰分析のイメージ

- **2つの説明変数による重回帰式のイメージ**
  - **重回帰モデル**（**データによる記述**）

$$y_i = b_0 + b_1 x_{i1} +, \cdots, +b_P x_{iP} + e_i \ \ (i = 1, \cdots N)$$



偏回帰係数の推定

- **最小2乗法による残差平方和の最小化**
  - **残差**　$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_{i1} +, \cdots, + b_P x_{iP})$
- **最小2乗推定量**

$$\begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \vdots \\ \hat{b}_P \end{bmatrix} = (X^T X)^{-1} X^T \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1P} \\ 1 & x_{21} & \cdots & x_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{NP} \end{bmatrix}$$

※上式の意味を理解するには線形代数の知識が必要。よって，線形代数を未学習の学生は，数学的な手続きによって回帰係数 $b_0, b_1, \cdots, b_P$ が推定可能ということを理解できれば十分

※現代では回帰分析はコンピューターを利用し，手作業で計算することはほぼ無い。後述のように，結果の意味を理解できることに授業の重点を置く

重回帰分析をしてみよう～問題設定

- **背景**
  - マーケッターとして，あるスーパーマーケットチェーンの販売促進に関する戦略立案を担うことになった
- **利用できるデータ**
  - ID-POSデータ：レジ通過時にポイントカードを提示した顧客の購買履歴データ。「誰が，いつ，何を，何個，いくらで」購入したのかが記録されているデータ（１年分を利用）
  - 顧客属性データ：ポイントカード登録顧客の年齢，家族人数，世帯内の高齢者の有無，世帯内の子供の有無，自宅から店舗までの所要時間
- **目的**
  - 顧客属性と購買金額の関係を定量的に把握するための因果型リサーチ

重回帰分析をしてみよう～データ#1

- ID-POSデータ

購買日	購買時間	顧客ID	商品カテゴリコード	商品コード	価格	購買個数
2020.05.15	11.15.01	100001	12321	490000000012	198	3
2020.05.15	11.15.01	100001	10089	490111234020	258	1
2020.05.15	11.16.11	123456	10105	490000675962	154	2
…	…	…	…	…	…	…

- 顧客属性データ

顧客ID	年齢	家族人数	高齢者の有無	子供の有無	家からの距離
00001	61	3	0	0	15分
00002	40	4	0	1	10分
00003	59	2	0	0	25分
…	…	…	…	…	…

- 重回帰分析用に加工した購買金額データ

顧客ID	購買金額	年齢	家族人数	高齢者の有無	子供の有無	家からの距離
00001	¥267,120	61	3	0	0	15分
00002	¥156,990	40	4	0	1	10分
00003	¥143,428	59	2	0	0	25分
…	…	…	…	…	…	…
01000	¥84,143	71	2	1	0	5分

重回帰分析をしてみよう～データ#2

- **重回帰分析用に加工した購買金額データ**
  - このデータがあれば目的の回帰分析が可能

顧客ID	購買金額	年齢	家族人数	高齢者の有無	子供の有無	家からの距離
00001	¥267,120	61	3	0	0	15分
00002	¥156,990	40	4	0	1	10分
00003	¥143,428	59	2	0	0	25分
…	…	…	…	…	…	…
01000	¥84,143	71	2	1	0	5分

～データと数式の対応表～

顧客ID	購買金額	年齢	家族人数	高齢者の有無	子供の有無	家からの距離
$i = 1$	y_1	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
$i = 2$	y_2	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}
$i = 3$	y_3	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}
…	…	…	…	…	…	…
$i = 1000$	y_N	x_{N1}	x_{N2}	x_{N3}	x_{N4}	x_{N5}

※本講義で利用するデータは実データを元に授業用に作成したダミーデータである。しかし、その分析結果は実際のデータの傾向が反映されている

重回帰分析をしてみよう～結果

- **重回帰分析の結果**

	Estimate (推定値)	Std.Error (標準誤差)	t value (t値)	Pr(>|t|) (p値)	
切片 (b_0)	106146	24196	4.39	0.000	***
年齢 (b_1)	841	382	2.21	0.028	*
家族人数 (b_2)	23170	2602	8.91	0.000	***
高齢者の有無 (b_3)	-1063	8202	-0.13	0.897	
子供の有無 (b_4)	7941	7633	1.04	0.299	
家からの時間 (b_5)	-3208	598	-5.37	0.000	**
Adjusted R-squared (自由度調整済み決定係数)		0.11			

  - ソフトウェアを利用することで, このような結果が出力される

- **これ以降の本授業の目標**
  ⇒ **この表の数値の意味を正しく解釈できる**