経済と社会 補助資料 ～重回帰分析の結果の解釈1～

2022年度1学期: 火曜2限
担当教員: 石垣 司

(再掲)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

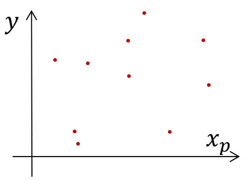
- これ以降の本授業の目標
  ⇒ この表の数値の意味を正しく解釈できる

回帰係数の検定

- 説明変数が目的変数に影響を与えるか否かの検定

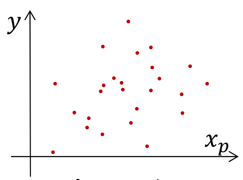
$$y = b_0 + b_1 x_1 +, \cdots, +b_P x_P$$

  - 説明変数 x_p が y に影響を与える: 回帰係数 $b_p \neq 0$
  - 説明変数 x_p が y に影響を与えない: 回帰係数 $b_p = 0$
    - 回帰係数の推定精度はサンプルサイズ n に依存

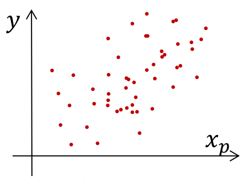


サンプルサイズ $n = 10$

$b_p \neq 0$ or $b_p = 0$???



サンプルサイズ $n = 25$

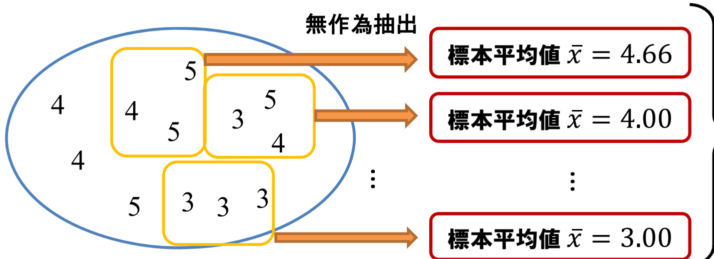


サンプルサイズ $n = 50$

$b_p \neq 0$ に思えるなぁ

標本分布

- 標本から計算される代表値(など)の確率分布
  - 例 $\mu = 4$ の母集団から $N = 3$ の標本抽出
    - 抽出される標本は確率的に決まる



母集団(母平均 $\mu = 4$)

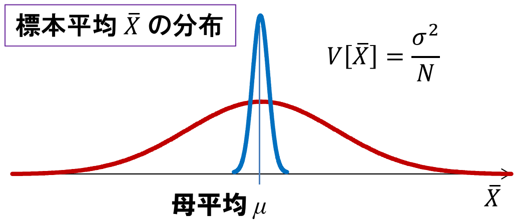
抽出された標本に依存して、実現値の値はばらつく

<u>標本平均値はある確率分布に従い発生する</u>

  - これ以降の設定 ⇒ 標本 $\{X_1, \ldots, X_N\}$ は<u>確率変数</u>
    - $\{x_1, \ldots, x_N\}$ は確率変数の実現値(観測値・データ)
    - 標本は独立同一分布から無作為抽出

標本平均の分布

- **標本平均** $\bar{X} = \frac{1}{n}\sum_{i=1}^{N} X_i$
  - $\bar{X}$ **は確率変数（サンプルサイズ** n **の標本抽出を1回の試行とみなすと, 試行ごとに標本平均の実現値は確率的に決まる）**
- **標本平均の期待値** $E[\bar{X}] = \mu$
- **標本平均の分散** $V[\bar{X}] = \frac{\sigma^2}{N}$

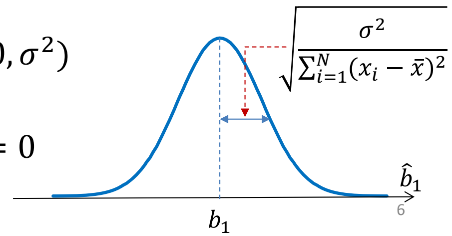


N **が大きいと分散が小さい**
$|\mu - \bar{x}|$ **が小さくなる確率が高い**

N **が小さいと分散が大きい**
$|\mu - \bar{x}|$ **が大きくなる確率が高い**

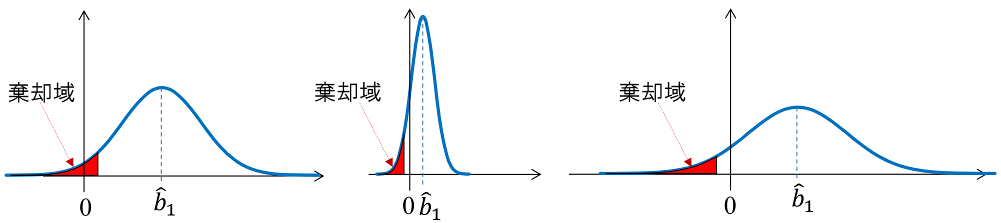
単回帰係数の検定 #1

- **単回帰モデル：** $y = b_0 + b_1 x_1$
  - **データ表現：** $y_i = b_0 + b_1 x_{i1} + e_i \ \ (i = 1, \cdots N)$
- **最小2乗推定量** $\hat{b}_1$ **は母集団からのサンプリングとサンプルサイズ** N **に依存する確率変数**
- **次の3つの仮定を満たすとき,** $\hat{b}_1 \sim N\left(b_1, \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right)$
  1. **説明変数** x_1 **は非確率的**
  2. **誤差項は正規分布：** $e_i \sim N(0, \sigma^2)$
     - **平均はゼロで不均一分散**
  3. **誤差項は独立：** $\mathrm{Cov}[e_i, e_j] = 0$



単回帰係数の検定 #2

- **帰無仮説** H_0 **：** $b_1 = 0$
- **対立仮説** H_1 **：** $b_1 \neq 0$
- **検定統計量：** $t = \frac{\hat{b}_1}{SE(\hat{b}_1)} \sim t_\alpha^{(n-2)}$
  - $SE(\hat{b}_1) = \sqrt{\frac{\sum_{i=1}^{N} e_i^2}{N-2} \frac{1}{\sum_{i=1}^{N}(x_i - \bar{x})^2}}$, $t_\alpha^{(n)}$ **:自由度** n **の** t **分布**



$b_1 = 0$**である確率は統計的に有意に小さい**

$b_1 = 0$**である確率は統計的に小さいとは言えない**

$b_1 = 0$**である確率は統計的に小さいとは言えない**

重回帰係数の検定

- **重回帰モデルの個別の回帰係数の検定**
  - **帰無仮説** H_0 **：** $b_p = 0$
  - **対立仮説** H_1 **：** $b_p \neq 0$
  - **検定統計量：** $t = \frac{\hat{b}_p}{SE(\hat{b}_p)} \sim t_\alpha^{(N-P-1)}$
    - $SE(\hat{b}_p) = \sqrt{\frac{\sum_{i=1}^{N} e_i^2}{N-P-1} \frac{1}{\sum_{i=1}^{N}(x_{ip} - \bar{x}_p)^2}}$
- **重回帰モデルの回帰係数の** F **検定**
  - **帰無仮説** H_0 **：** $b_1 =, \ldots, = b_P = 0$
  - **対立仮説** H_1 **：** b_1 **から** b_P **のどれかがゼロではない**

回帰係数の検定とP 値

- **観測された有意水準**
  - P **値が事前に設定された有意水準** α **よりも小さければ，帰無仮説は棄却できる**
  - **左側検定での例**
    - **統計量** $t = \frac{\hat{b}_p}{SE(\hat{b}_p)} \sim t_\alpha^{(N-P-2)}$
    - **例：** $t = -2.5$**のとき**

$$p \text{ value} = P(t < -2.5 | b_1 = 0)$$

  - **コンピュータによる計算から** P **値** $= 0.027$
  - 5%**水準では有意差あり**，1%**水準では有意差なし**

重回帰分析の結果の解釈 #1

	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		

- **赤字部分の解釈**

「年齢(b_1)，家族人数(b_2)，家からの距離(b_5)は，有意水準5%で統計的に有意に購買金額に影響を与えている」

  - **高齢者の有無(b_3)，子供の有無(b_4)は購買金額に影響を与えているかどうかは分からない**

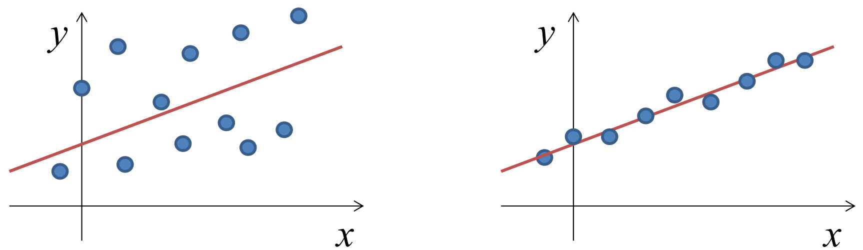
重回帰分析の結果の解釈 #2

	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		

- **誤った赤字部分の解釈**
  - **高齢者の有無(b_3)，子供の有無(b_4)は購買金額に影響を与えていない**
  - **家族人数(b_2)のP値が一番低いので，家族人数が最も強く購買金額に影響を与えている**
  - **有意水準を1%に設定すると，年齢(b_1) が有意ではなるので有意水準5%の方が良い分析結果である**

決定係数 R^2

- **回帰式の適合度**(goodness of fit)**の指標**
  - **同じ回帰式でもデータの説明力が異なる**



説明力が低い回帰直線 ($R^2 \approx 0$)　**説明力が高い回帰直線** ($R^2 \approx 1$)

異なるデータから推定された同じ回帰直線

- **決定係数** R^2　$(0 \leq R^2 \leq 1)$
  - **適合度が高いと**1**に近く，低いとゼロに近い**

決定係数 R^2 の定義と意味

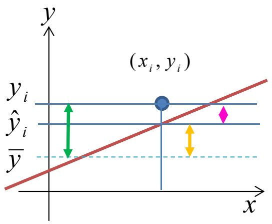
- **決定係数 R^2 の定義**

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

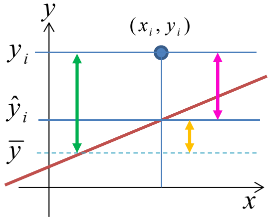
全変動(TSS: total sum of squares)
回帰変動(ESS: explained SS)
残差変動(RSS: residual SS)

  - **標本分散の分解**

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{TSS}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\text{ESS}} + \underbrace{\sum_{i=1}^{n}e_i^2}_{\text{RSS}}$$



適合度が高いデータの例

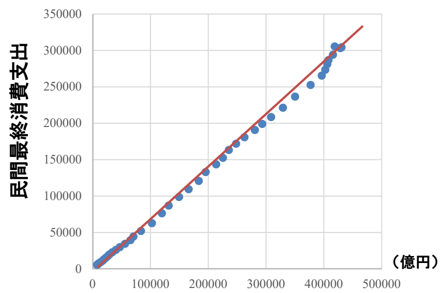


適合度が低いデータの例

$y_i - \bar{y}$ TSSの一部
$\hat{y}_i - \bar{y}$ ESSの一部
e_i RSSの一部

決定係数の例～所得と消費

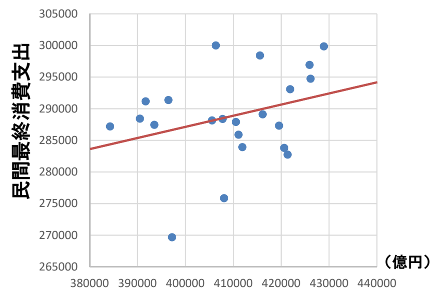
- **国民可処分所得と民間最終消費支出**（内閣府　国民経済計算年次推計）



国民可処分所得
1955年度～1998年度(1968SNA)

$\hat{a} = -2950,\ \hat{b} = 0.698,\ R^2 = 0.998$

消費額の全変動の99.8%は
国民可処分所得で説明可能



国民可処分所得
1994年度～2015年度(2008SNA)

$\hat{a} = 215600,\ \hat{b} = 0.178,\ R^2 = 0.099$

$\hat{a}$ 基礎消費、
$\hat{b}$ 限界消費性向

消費額の全変動の約10%は
国民可処分所得で説明可能

重回帰分析と決定係数

- **自由度調整済み決定係数 $\tilde{R}^2$**

$$\tilde{R}^2 = \frac{(N-1)R^2 - P}{N-P-1}$$

  - 重回帰分析では説明変数の数 P を大きくすると, 残差2乗和は単調減少する ⇒ 決定係数の値が大きくなる
  - $\tilde{R}^2$ は R^2 を N と P の値で調整している
  - 重回帰分析の当てはまりの良さをみるためには自由度調整済み決定係数 $\tilde{R}^2$ を用いるのが一般的

重回帰分析の結果の解釈 #1

	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		

- **赤字部分の解釈**

「回帰変動の大きさは全変動の大きさの約11%である」

  - ただし, あくまであてはまり具合の目安。モデル選択などに利用することの合理性はない