統計学入門
～単回帰分析～

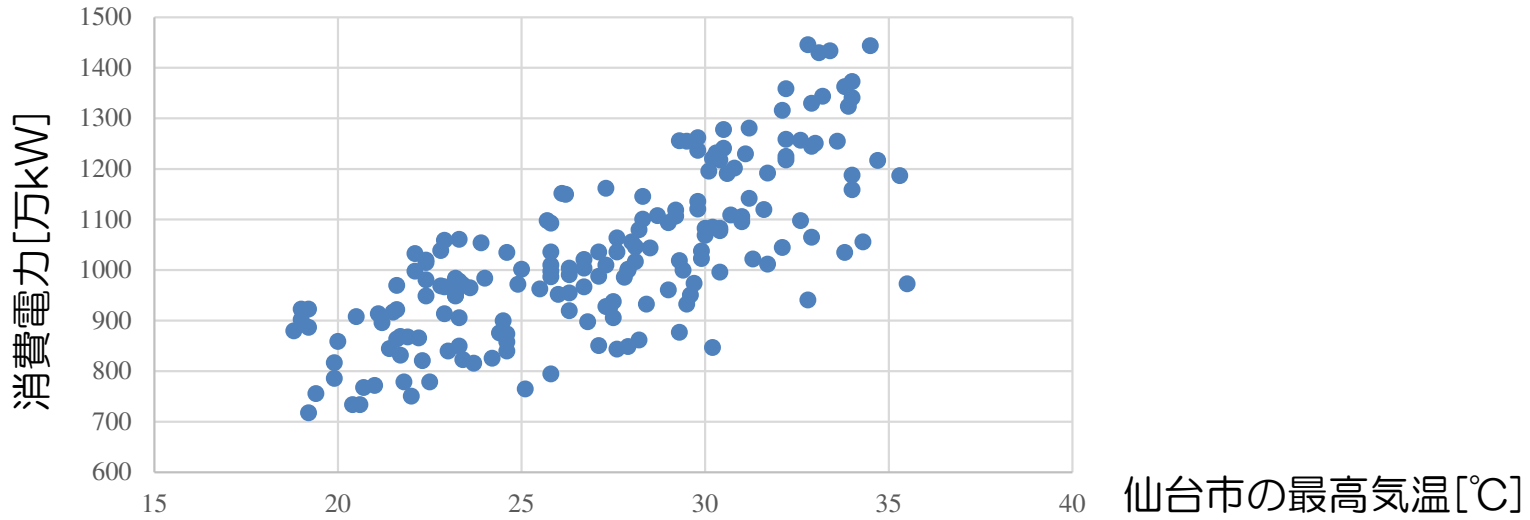
2025年度1学期:　　月曜2限

担当教員:　　石垣　司

回帰分析の前に

基本は「散布図」

- 変数 x（最高気温）と y（消費電力）の相関関係の可視化
  - この関係性を利用した予測や実証の手段が回帰分析



「回帰」とは，目的変数 y の動きを，別の説明変数 x と関数 f で予測したり説明したりすること

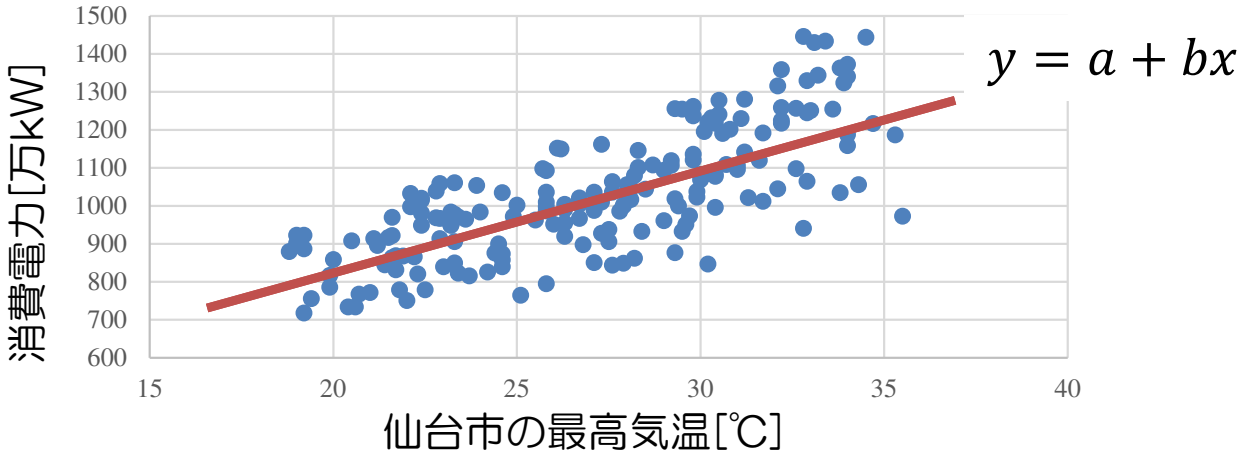
東北電力ネットワーク：東北6県・新潟エリアの2020&21年7月1日～9月30日の各日の12時から13時の電力使用量[万kW]
https://setsuden.nw.tohoku-epco.co.jp/download.html
気象庁：2020&21年7月1日～9月30日の各日の仙台市の最高気温
https://www.data.jma.go.jp/obd/stats/etrn/

線形単回帰モデル

関数 f に直線を仮定した説明変数が1つのときの回帰分析のためのモデル

$$y = f(x) = a + bx$$

Practice!



- **因果関係がある場合は, x が原因, y が結果の表現**
- **データ$\{(x_1, y_1), \cdots, (x_N, y_N)\}$を用いて, 散布図の傾向に適合する直線の切片 a と傾き b を推定**
- **切片 a と傾き b が決まれば, 目的変数 y を予測できる**

単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

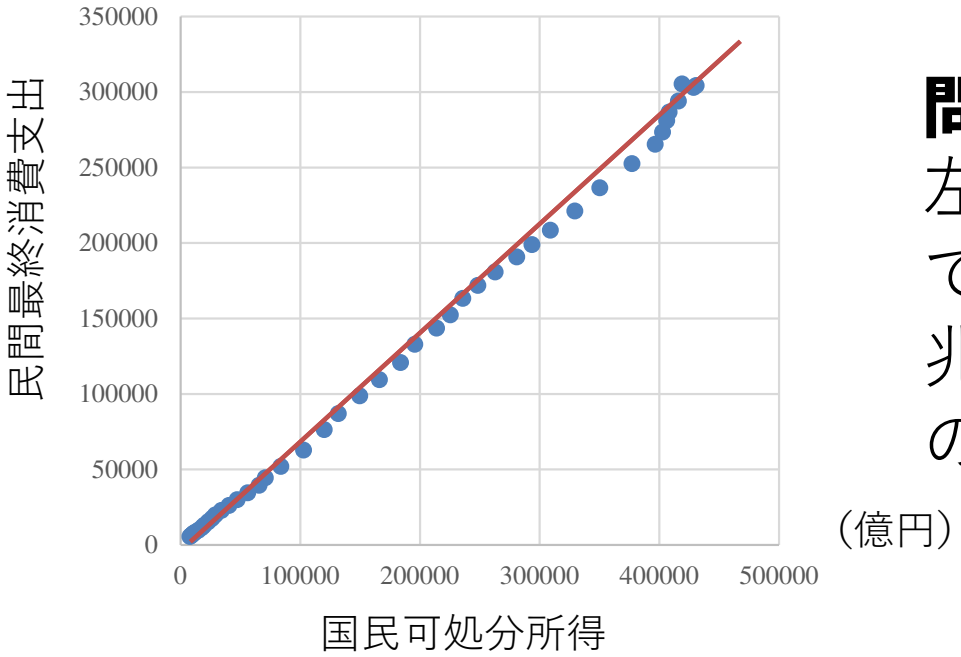
回帰分析と予測

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？

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**

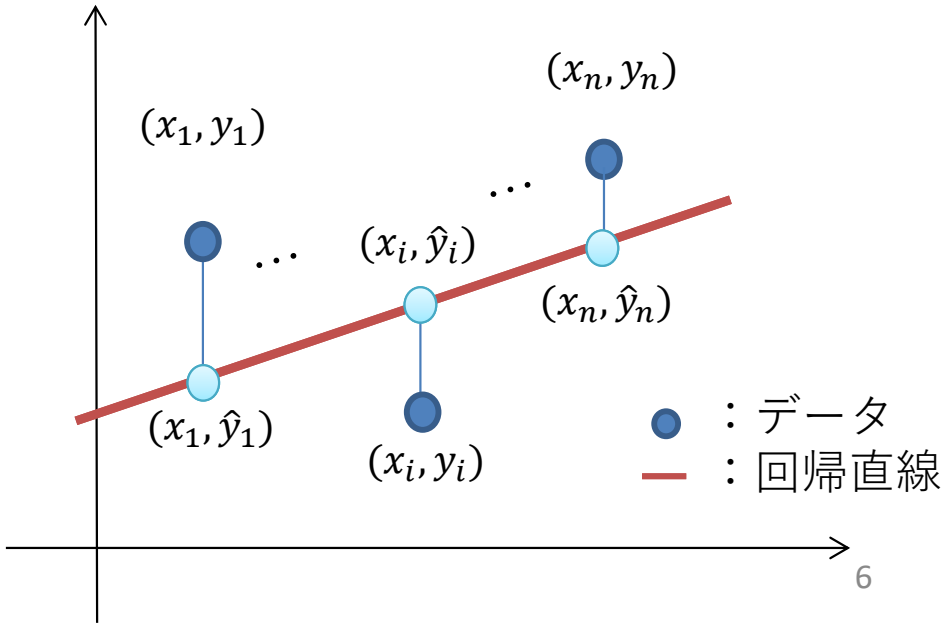
基準は残差平方和(RSS: Residual sum of squares)**の最小化**

- **残差**e_i　　$e_i = y_i - \hat{y}_i = y_i - (a + bx_i)$
- **残差平方和**　　$RSS = \sum_{i=n}^{n} e_i^2$

手続きは最小2乗法

最小2乗推定量　check!

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

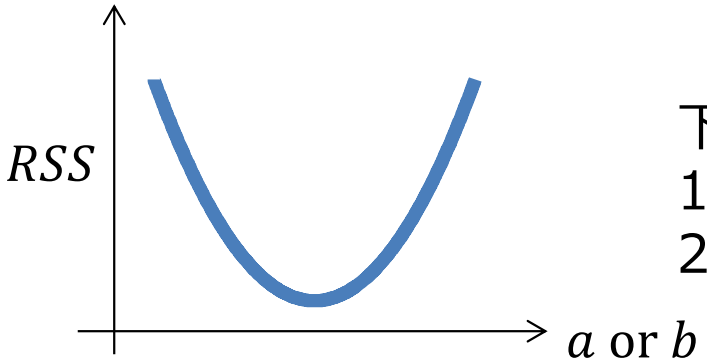


最小2乗法

データとの誤差が全体的に小さくなる関数の推定法

- **正則な場合，合理的な手続きで解が一つに定まる**
- **残差平方和(RSS)は変数 a と b に関する2次関数**

$$RSS = f(a,b) = \sum_{i=n}^{n} e_i^2 = \sum_{i=n}^{n} \{y_i - (a + bx_i)\}^2$$



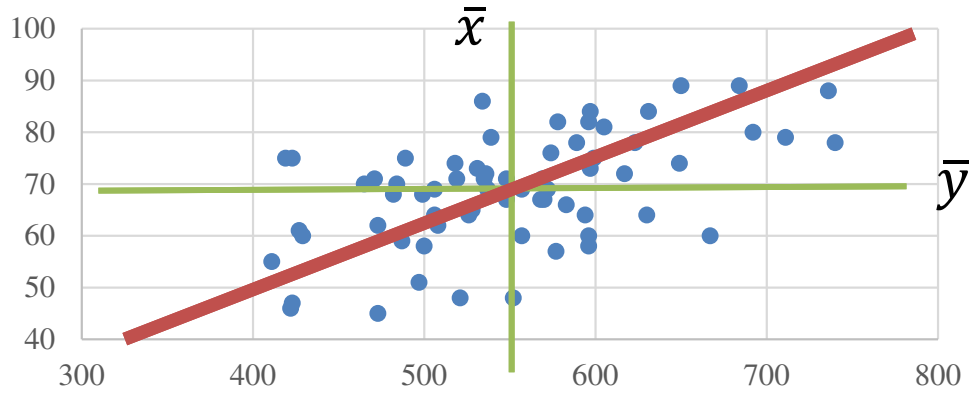
下に凸な2次関数の特徴
1. 必ず最小点がある
2. 最小点の傾きは 0

- **RSS が最小となる $\hat{a}$ と $\hat{b}$ は下式を満たす**

$$\frac{\partial f(a,b)}{\partial a} = 0, \frac{\partial f(a,b)}{\partial b} = 0$$

推定された回帰式が満たす性質

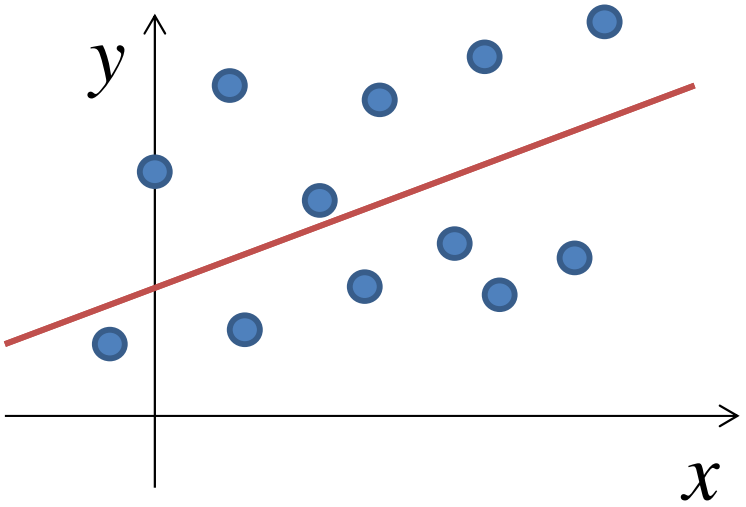
1. **推定された回帰直線は$(\bar{x}, \bar{y})$を通る** check!
2. $\sum_{i=1}^{n} e_i = 0$（**残差の和は0**） check!
3. $\hat{b} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{S_{xy}}{S_{xx}}$ check!
4. $\sum_{i=1}^{n} e_i x_i = 0$（**残差と説明変数 x の積和は0**） check!
   - **残差と説明変数のベクトルは直交する**



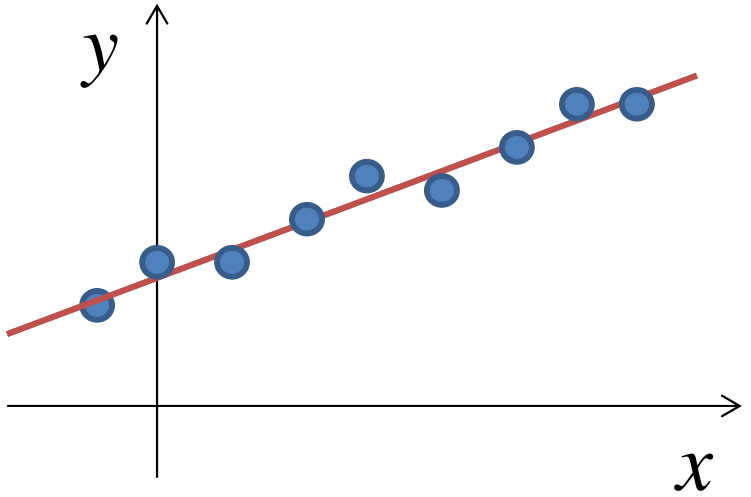
決定係数 R^2

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 $\quad (0 \leq R^2 \leq 1)$

– 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

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

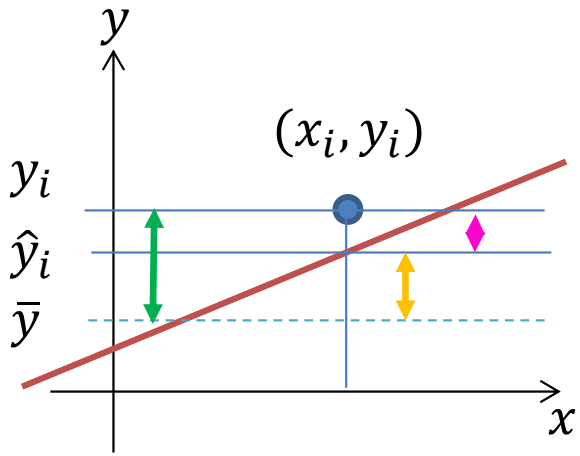
全変動(TSS: total sum of squares)
回帰変動(ESS: explained SS)
残差変動(RSS: residual SS)

– 標本分散の分解

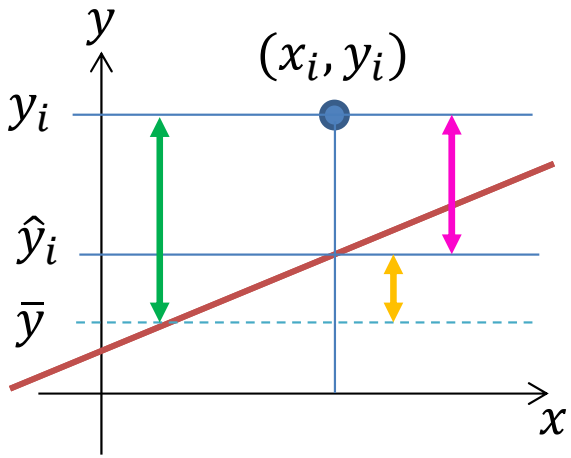
check!

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2$$

TSS　　ESS　　RSS



適合度が高いデータの例



適合度が低いデータの例

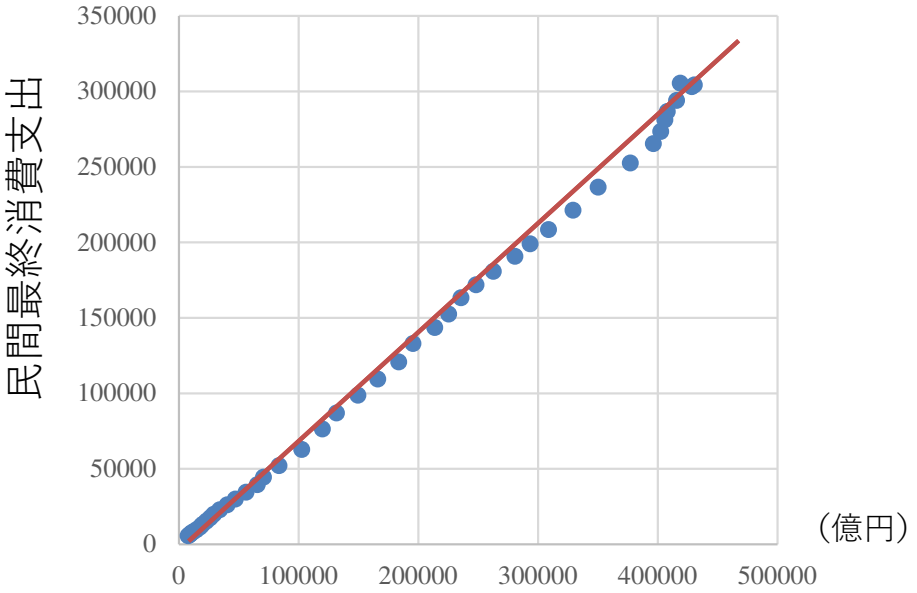
$y_i - \bar{y}$ TSSの一部

$\hat{y}_i - \bar{y}$ 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）

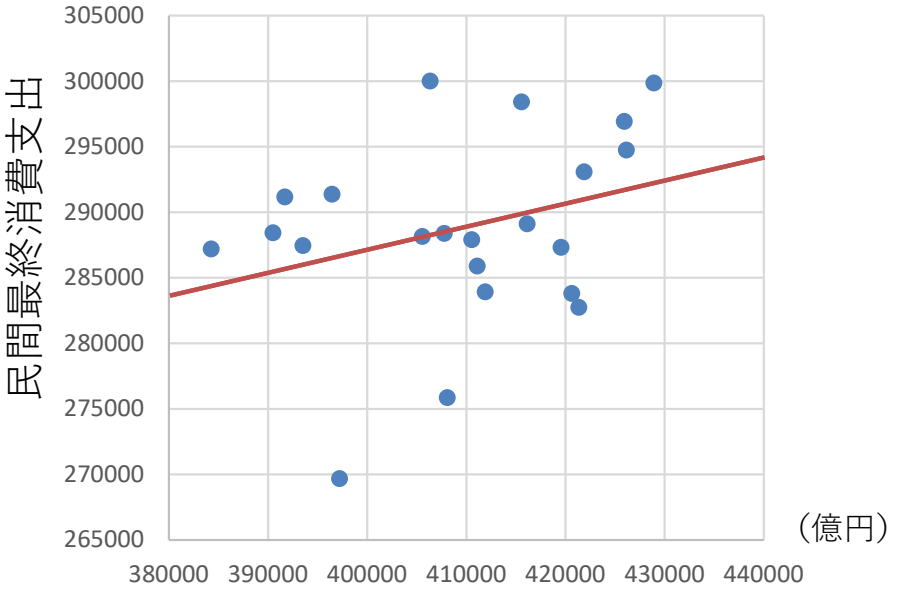


国民可処分所得

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)

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

消費額の全変動の約10%は国民可処分所得で説明可能

$\hat{a}$	基礎消費、
$\hat{b}$	限界消費性向

補足：偏相関係数～計算方法

標本偏相関係数の定義

- 回帰により z の影響を除去した x の残差

$$u_i = x_i - (\hat{c} + \hat{d}z_i)$$

- 回帰により z の影響を除去した y の残差

$$v_i = y_i - (\hat{e} + \hat{f}z_i)$$

- z の影響を除去した x と y の標本偏相関係数

$$r_{xy,z} = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} u_i^2 \sum_{i=1}^{n} v_i^2}} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}$$

演習問題

ある商品に関する広告費(x), 売上数(y), 価格(z)の各データの代表値は次であった

- **標本平均** $\bar{x} = 80, \bar{y} = 100, \bar{z} = 50$
- **標本分散** $S_{xx} = 16, S_{yy} = 16, S_{zz} = 1$
- **標本共分散** $S_{xy} = 12, S_{xz} = -2, S_{yz} = -3$
- **標本相関係数** $r_{xz} = -0.5, r_{yz} = -0.75$

問題

1. x **を説明変数,** y **を目的変数とした単回帰分析により, 広告費が** 120 **の時の売上数を予測しなさい**
2. **相関係数** r_{xy} **の値を求めなさい**
3. z **の影響を削除した偏相関係数** $r_{xy,z}$ **の値を求めなさい**