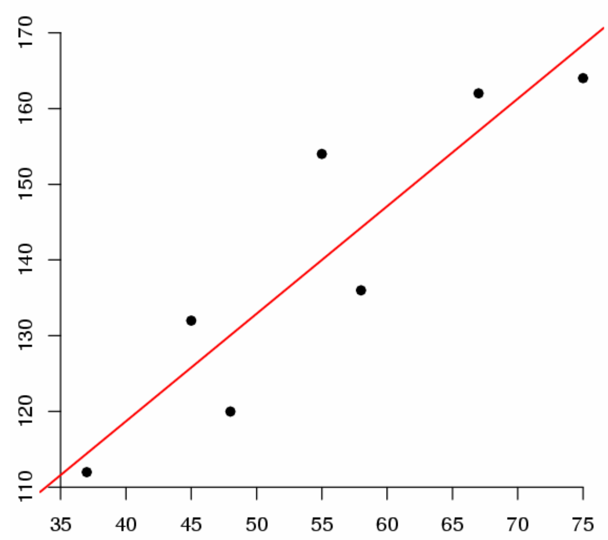
回帰直線

回帰直線

- 相関係数：変量間の直線的関連性の強さを測る尺度
- 2つの変量間の関係を直線で表現する方法を考える
  - 測定の難しい変量を，測定が容易な変量を用いて推測する
  - ある変量を，他の変量から予測する
- 基本的な考え方
  - 2つの変量に直線的な関係を仮定し，直線の傾きと切片を推定する
- 右図
  - データ：年齢（横軸）と血圧値（縦軸）
  - 赤い直線： $y = 63 + 1.4x$

回帰直線

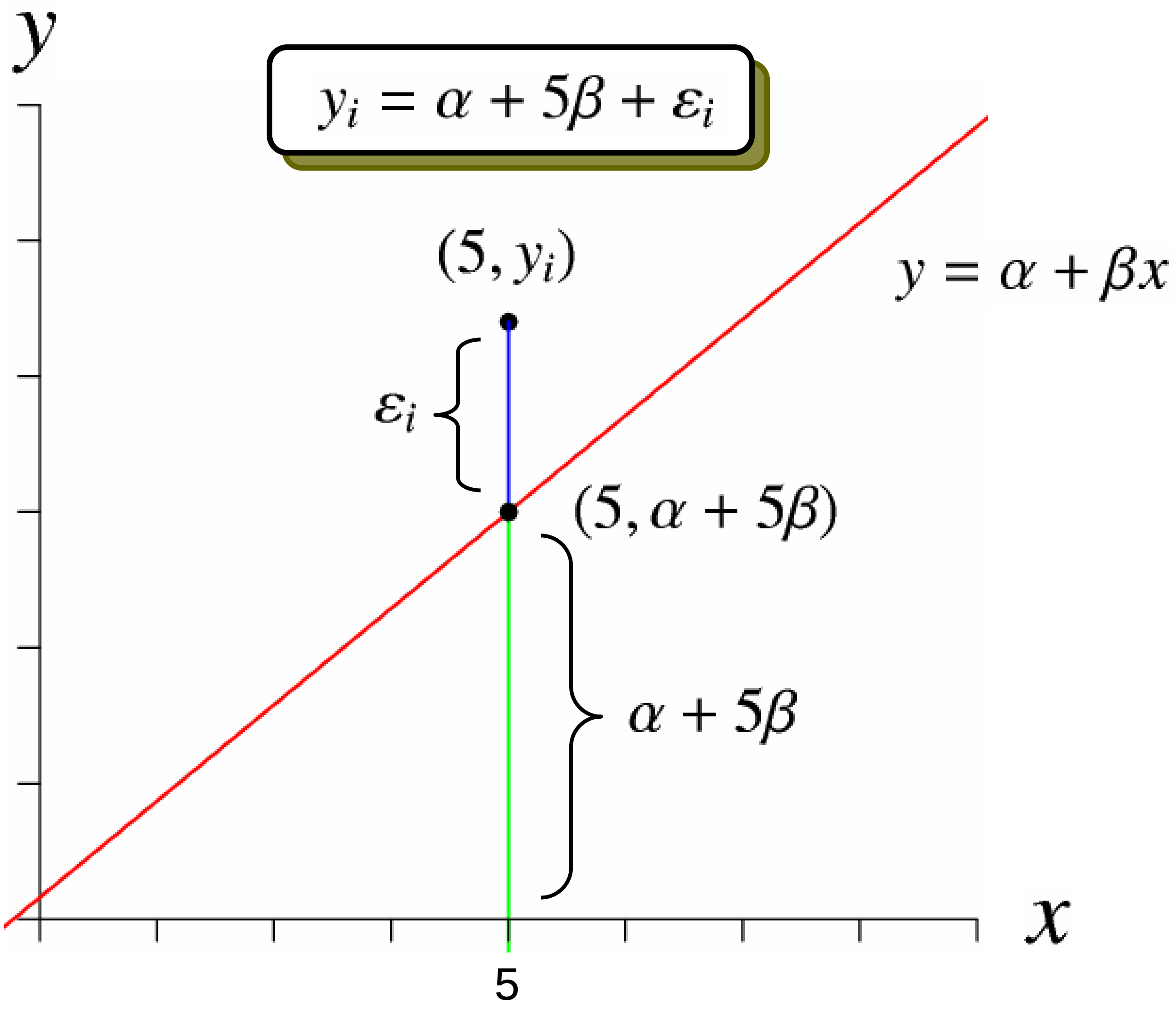
$$y = 63 + 1.4x$$



回帰モデル

- 2つの変量の間に直線関係を仮定することが妥当な場合，次のような回帰モデルを考えられる
- 回帰モデル
  - $y = \alpha + \beta x + \varepsilon$
  - ε : 誤差項
  - $\alpha + \beta x$ を平均として, y はその近くの値をとる
- 問題点
  - α と β の決め方

回帰モデルの構成



y
$y_i = \alpha + 5\beta + \varepsilon_i$
$(5, y_i)$
$y = \alpha + \beta x$
ε_i
$(5, \alpha + 5\beta)$
$\alpha + 5\beta$
x
5

誤差項について

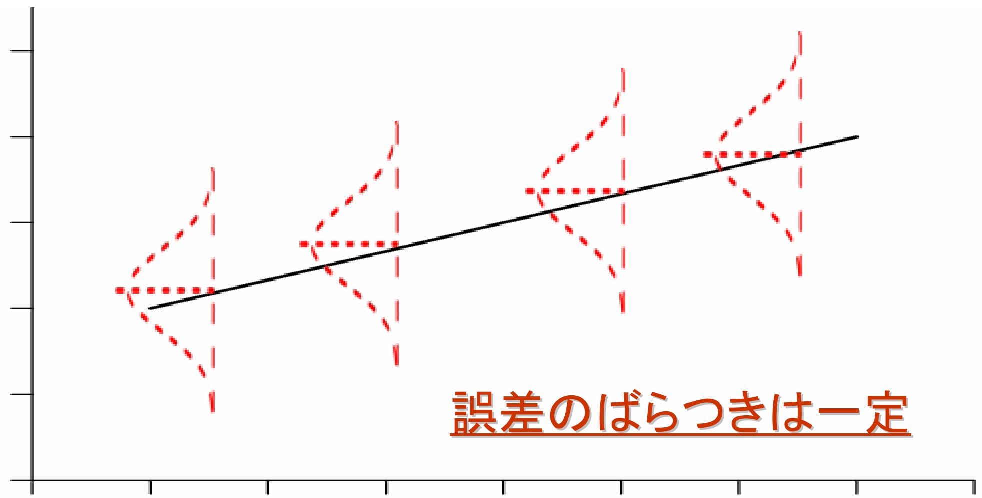
- **誤差項ε**
  - モデルに取り込んだ変数のみでは説明しきれない情報
  - 測定誤差

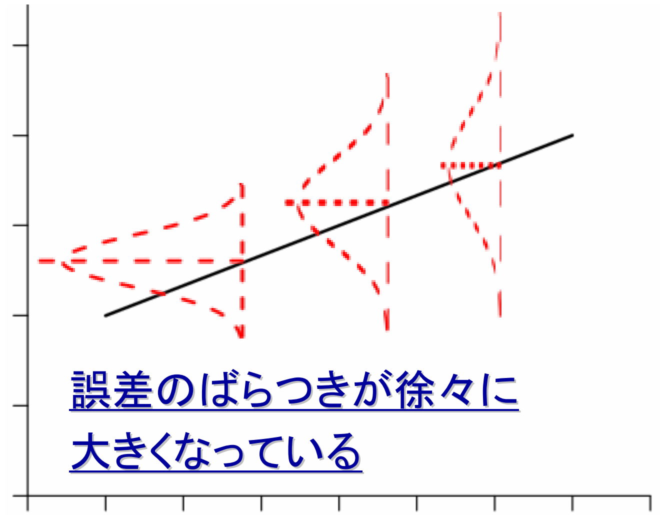
  等

- **誤差に関する3つの条件**
  - ε_iとε_jとは独立である $(i \neq j,\ i, j = 1, 2, \ldots, n)$
  - ε_iの平均は0である $(i = 1, 2, \ldots, n)$
  - ε_iの分散は全て等しい $(i = 1, 2, \ldots, n)$

誤差項の条件について

- ε_i の平均は 0 である $(i = 1, 2, \ldots, n)$
  - 目的変数の平均が回帰直線上にある
- ε_i の分散は全て等しい $(i = 1, 2, \ldots, n)$
  - 左下の図：条件を満たす場合
  - 右下の図：条件を満たさない場合

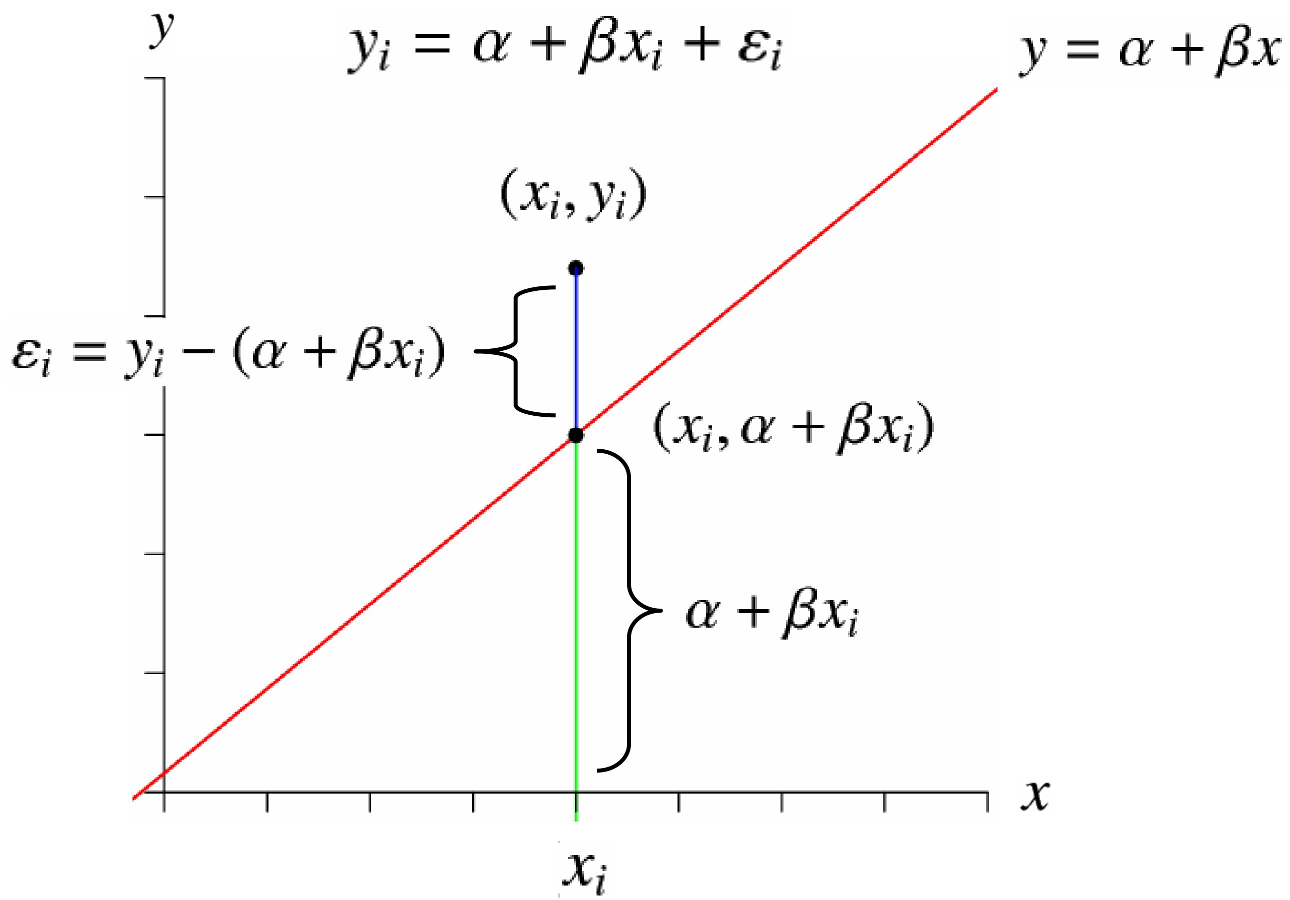




回帰係数の推定

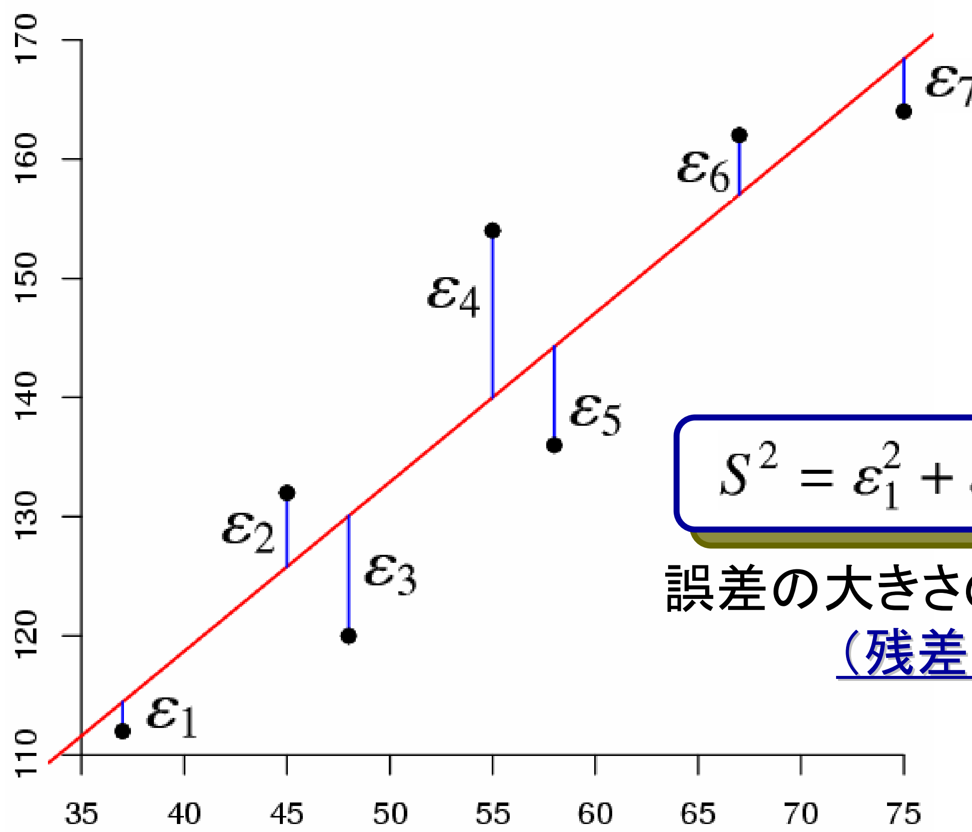
基本的な考え方

ε_i の二乗和が最小となるように α と β を定める



誤差の大きさ

ε_i の二乗和が最小となるように α と β を定める



$$S^2 = \varepsilon_1^2 + \varepsilon_2^2 + \cdots + \varepsilon_n^2$$

誤差の大きさの総合的な指標
（残差平方和）

回帰係数の推定量

- ε_i の二乗和

$$\begin{aligned} S^2 &= \varepsilon_1^2 + \varepsilon_2^2 + \cdots + \varepsilon_n^2 \\ &= \{y_1 - (\alpha + \beta x_1)\}^2 + \{y_2 - (\alpha + \beta x_2)\}^2 + \cdots + \{y_n - (\alpha + \beta x_n)\}^2 \end{aligned}$$

- S^2 を最小にするような回帰モデルは次のように表される
  - $\hat{y} = \hat{\alpha} + \hat{\beta} x$
  - $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$
  - $\hat{\beta} = \dfrac{(x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}$

例：回帰係数の推定

- 目的
  - 年齢から血圧を推測するための回帰モデルを構成する
- データ
  - 7名の女性の年齢を調査し，心臓収縮血圧を測定

被験者	1	2	3	4	5	6	7
年齢 x	37	45	48	55	58	67	75
血圧 y	112	132	120	154	136	162	164

- 回帰モデル
  - $y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, 2, \ldots, 7$

回帰係数の推定

- 各変数の平均値
  - $$\bar{x} = \frac{37 + 45 + \cdots + 75}{7} = 55$$
  - $$\bar{y} = \frac{112 + 132 + \cdots + 164}{7} = 140$$
- 回帰係数の推定量
  - $$\hat{\beta} = \frac{(37 - 55)(112 - 140) + \cdots + (75 - 55)(164 - 140)}{(37 - 55)^2 + \cdots + (75 - 55)^2} = \frac{(-18)(-28) + \cdots + 20 \cdot 24}{18^2 + \cdots + 20^2} = 1.4$$
  - $$\hat{\alpha} = 140 - 1.4 \times 55 = 63$$

推定された回帰モデル

- 推定された回帰モデル
  - $\hat{y} = 63 + 1.4x$
- 実現値・予測値・残差

被験者	年齢	血圧値	予測値	残差
1	37	112	114.5	-2.47
2	45	132	125.8	6.19
3	48	120	130.1	-10.07
4	55	154	140.0	14.00
5	58	136	144.3	-8.26
6	67	162	157.0	4.97
7	75	164	168.4	-4.38