データ解析 (第8回)

静岡大学システム工学科
安藤 和敏

第4章 単回帰分析

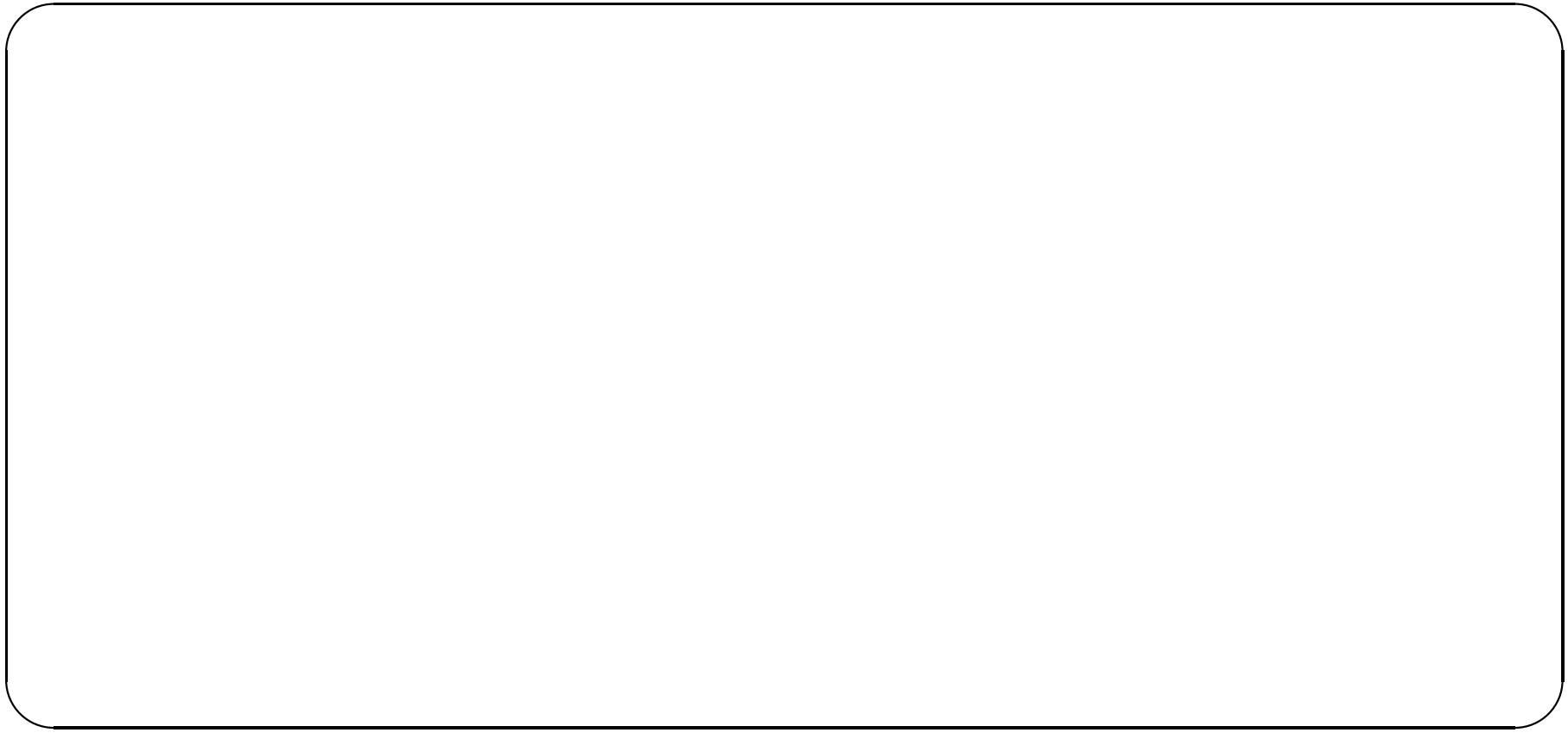
回帰分析法

浜松駅周辺の中古マンションのデータ

データ No.	広さ x_1 (m^2)	築年数 x_2 (年)	価格 y (千万円)
1	51	16	3.0
2	38	4	3.2
3	57	16	3.3
4	51	11	3.9
5	53	4	4.4
6	77	22	4.5
7	63	5	4.5
8	69	5	5.4
9	72	2	5.4
10	73	1	6.0

回帰分析によって明かになること

このデータについて, 以下の事を知りたいとする.



回帰分析によって明かになること

このデータについて, 以下の事を知りたいとする.

1. 価格は広さと築年数によって, 予測できるか.

回帰分析によって明かになること

このデータについて, 以下の事を知りたいとする.

1. 価格は広さと築年数によって, 予測できるか.
2. 予測できるとすれば, その精度はどれくらいか.

回帰分析によって明かになること

このデータについて, 以下の事を知りたいとする.

1. 価格は広さと築年数によって, 予測できるか.
2. 予測できるとすれば, その精度はどれくらいか.
3. 同じ地区で広さ $70m^2$, 築年数 10 年, 価格 5.8 千万円のマンションを提示された. この価格は妥当か.

回帰分析によって明らかになること

回帰分析法によって, 以下の事が分かる.

回帰分析によって明らかになること

回帰分析法によって, 以下の事が分かる.

1. 価格と広さと築年数は以下の関係にあると推定される.

$$y = 1.02 + 0.0668x_1 - 0.0808x_2$$

回帰分析によって明らかになること

回帰分析法によって, 以下の事が分かる.

1. 価格と広さと築年数は以下の関係にあると推定される.

$$y = 1.02 + 0.0668x_1 - 0.0808x_2$$

2. 寄与率は 0.933 で上式の精度は十分高い.

回帰分析によって明らかになること

回帰分析法によって, 以下の事が分かる.

1. 価格と広さと築年数は以下の関係にあると推定される.

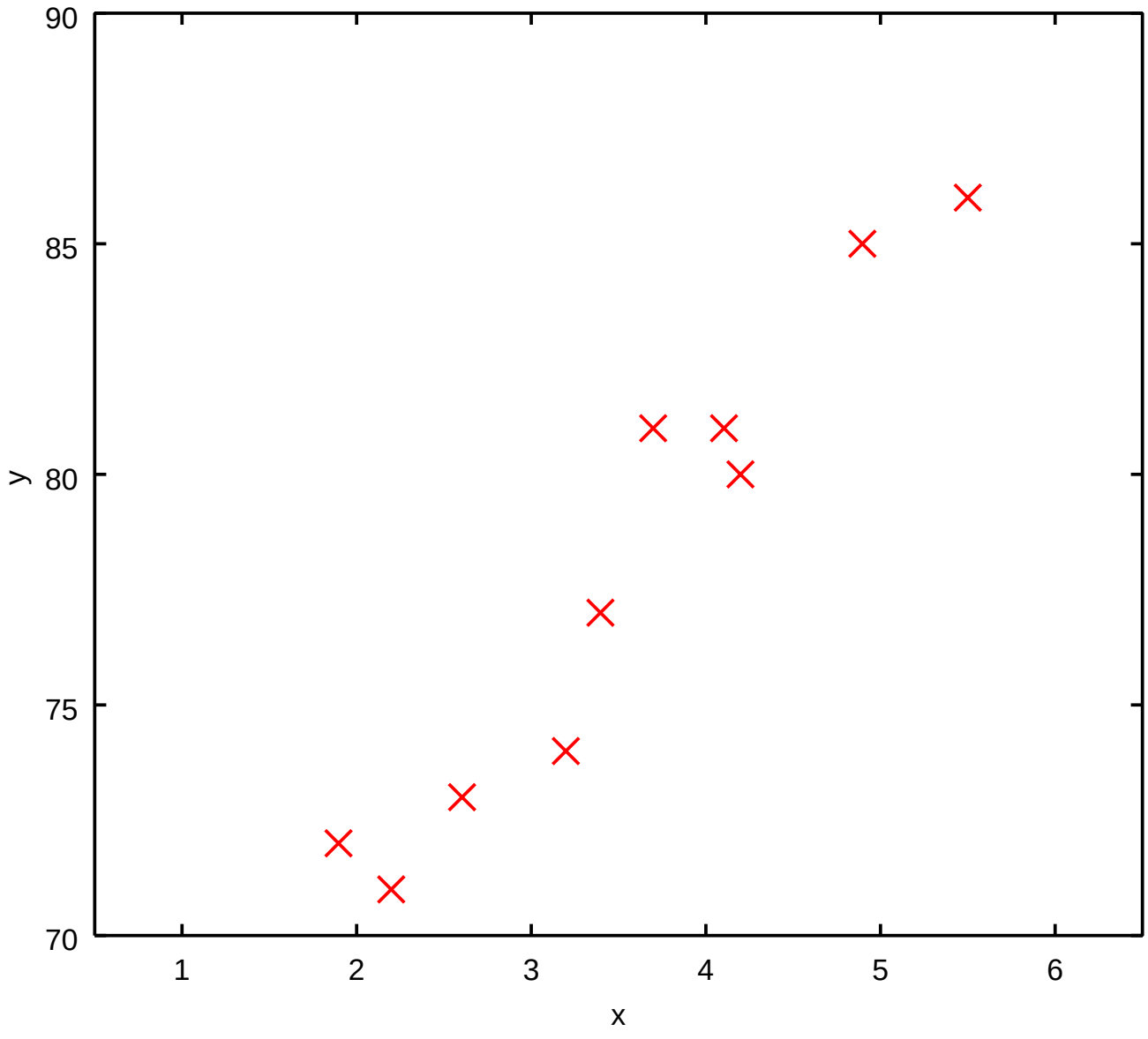
$$y = 1.02 + 0.0668x_1 - 0.0808x_2$$

2. 寄与率は 0.933 で上式の精度は十分高い.
3. $x_1 = 70, x_2 = 10$ を代入すると, $y = 4.89$ となるので, 5.8 千万円は相場より高い.

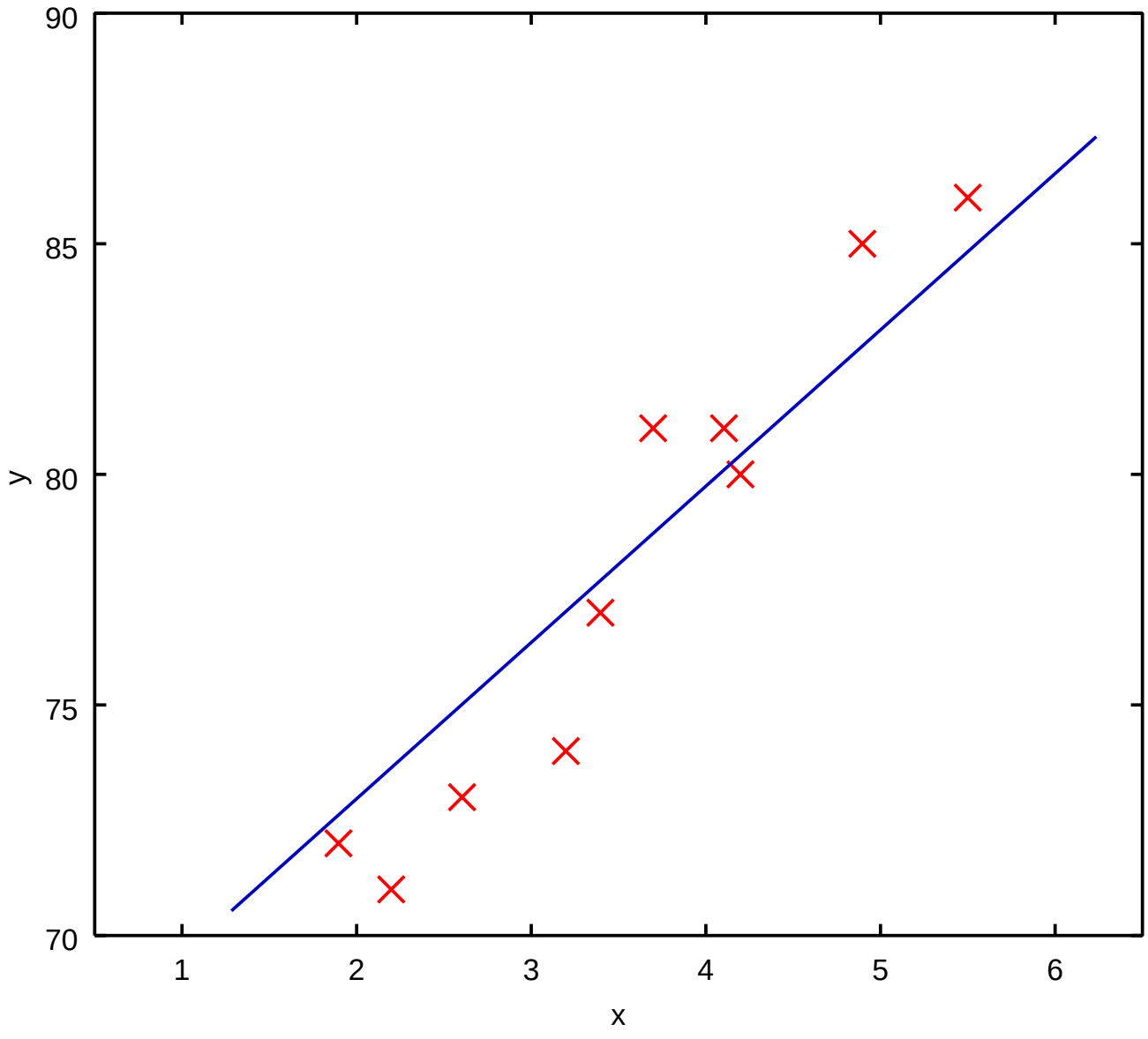
単回帰分析のデータの例(表4.1)

データ No.	含有量 x	収率 y
1	2.2	71
2	4.1	81
3	5.5	86
4	1.9	72
5	3.4	77
6	2.6	73
7	4.2	80
8	3.7	81
9	4.9	85
10	3.2	74

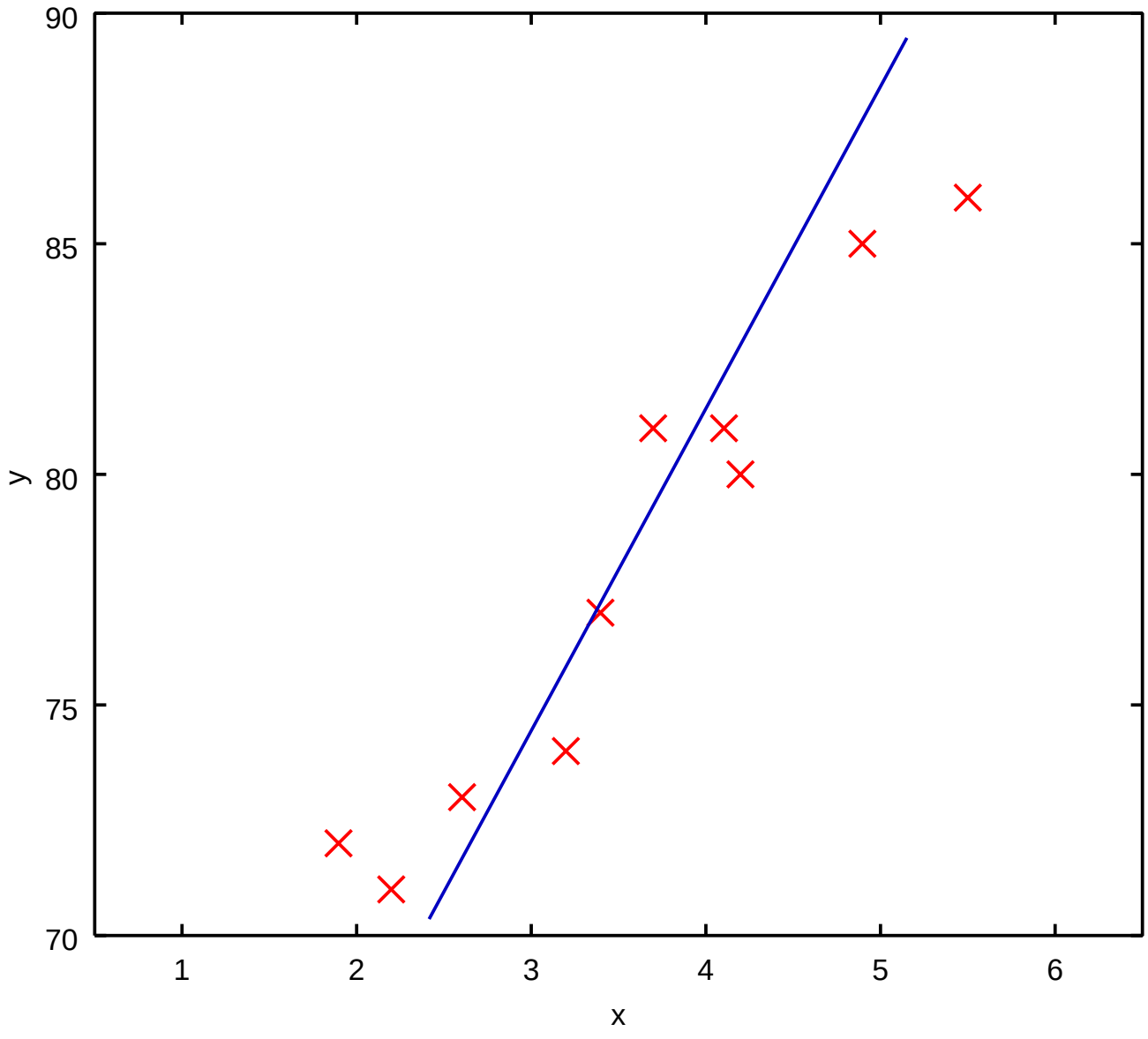
散布図



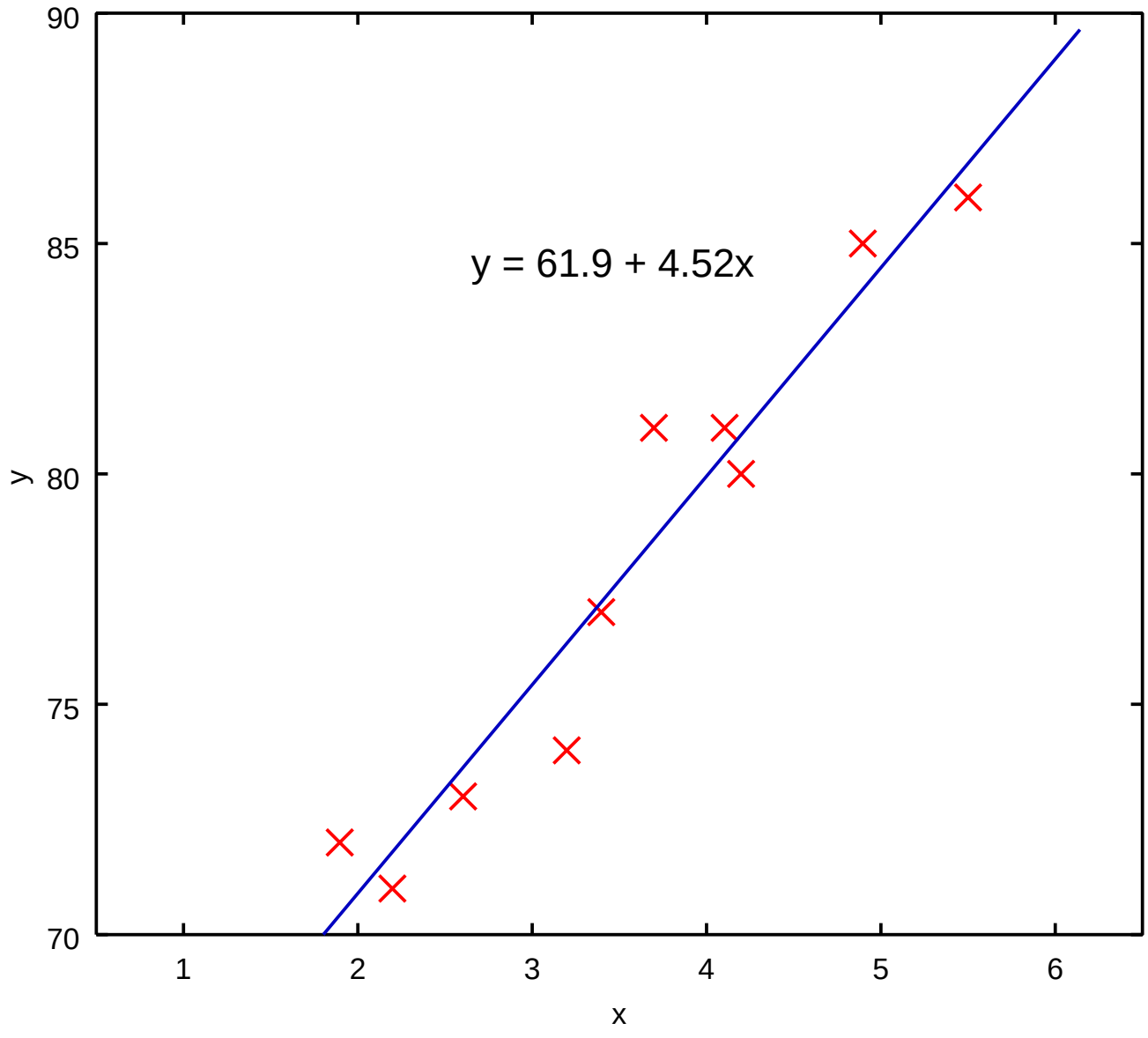
直線のあてはめ(1)



直線のあてはめ(2)



直線のあてはめ(3)



単回帰分析の目的の一つ

与えたれたデータに「最も良くあてはまる」
直線

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

を求めること. (つまり, y 切片 $\hat{\beta}_0$ と傾き $\hat{\beta}_1$ を求めること.)
ただし,「最も良くあてはまる」ということはどういうことなのかを数学的に定義しなければならない.

単回帰分析のデータ

データ No.	説明変数 x	目的変数 y
1	x_1	y_1
2	x_2	y_2
$\vdots$	$\vdots$	$\vdots$
i	x_i	y_i
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

直線のあてはまりの良さ

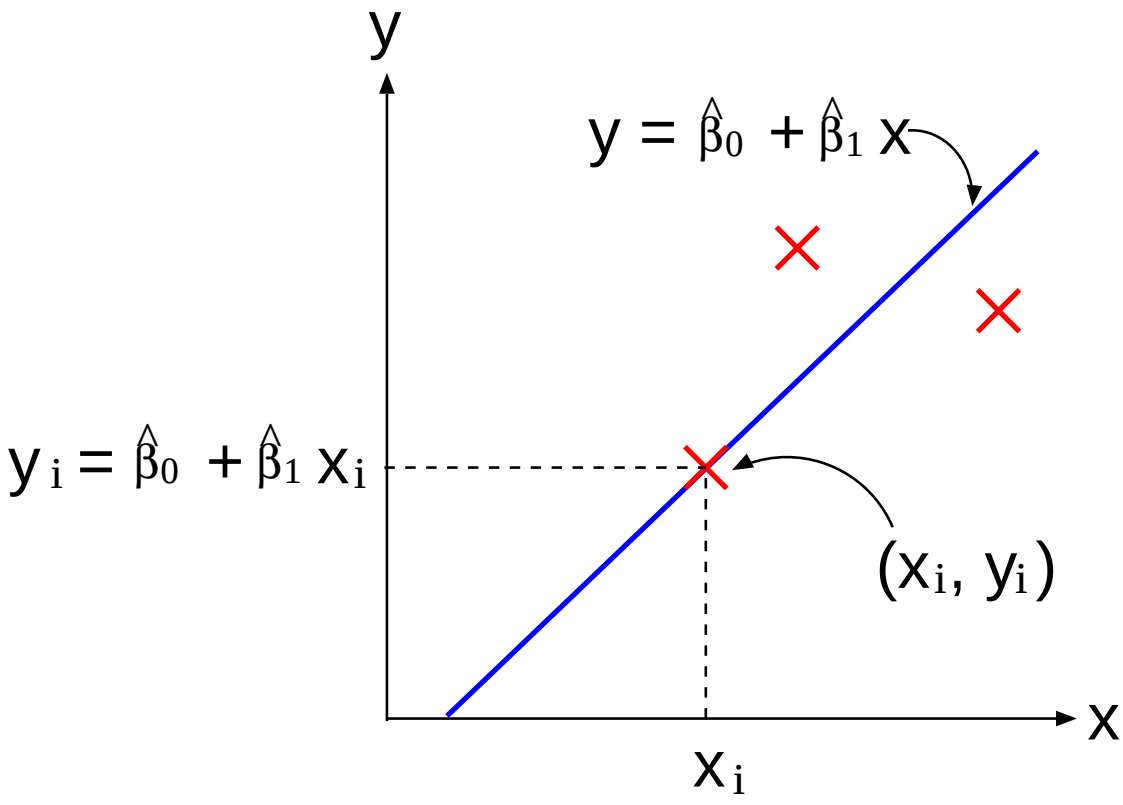
一つの直線

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

を固定して, この直線がどれだけこのデータにあてはまっているかを考える.

誤差

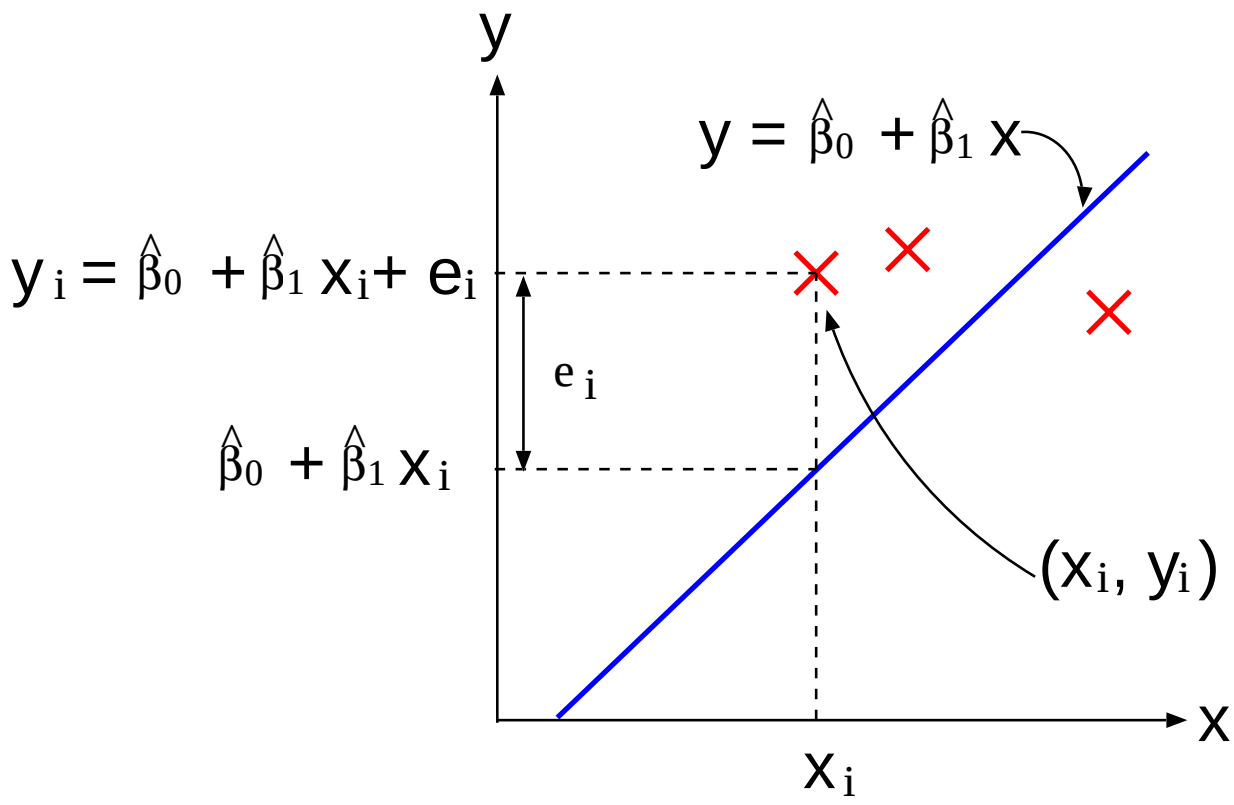
もし, 第i番目のデータ(x_i, y_i)が直線$y = \hat{\beta}_0 + \hat{\beta}_1 x$上にあれば, $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$であるが,



誤差

実際には誤差 e_i が加わって,

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i.$$



残差平方和

e_i は負にも正にもなるので, e_i^2 を第 i データの直線 $y = \hat{\beta}_0 + \hat{\beta}_1 x$ からの「ずれ」と考える. これらの「ずれ」を全てのデータについて足し合わせた量

$$S_e = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \{y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\}^2$$

を残差平方和と呼び, これが直線 $y = \hat{\beta}_0 + \hat{\beta}_1 x$ のデータへのあてはまりの程度を表していると考える.

最小2乗法

したがって, データに「最もよくあてはまる」直線$y = \hat{\beta}_0 + \hat{\beta}_1 x$を求める問題は,

$$S_e = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \{y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\}^2$$

を最小にするような, $\hat{\beta}_0$と$\hat{\beta}_1$を求める問題になった.

このようにして, $\hat{\beta}_0$, $\hat{\beta}_1$を求める方法を最小２乗法と呼ぶ.

思い出そう (2.2)式と(2.8)式

$$\text{平方和:}\quad S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$\text{偏差積和:}\quad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$\hat{\beta}_0$ と $\hat{\beta}_1$ の求めかた

$$S_e = \sum_{i=1}^{n} \{y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\}^2$$

は $\hat{\beta}_0$ と $\hat{\beta}_1$ を変数としてもつ関数である.
S_e の最小値を与える $(\hat{\beta}_0, \hat{\beta}_1)$ においては, S_e の偏微分が 0 となる.

$$\frac{\partial S_e}{\partial \hat{\beta}_0} = 0, \quad \frac{\partial S_e}{\partial \hat{\beta}_1} = 0.$$

$\hat{\beta}_0$ と $\hat{\beta}_1$ の求めかた

ゆえに, この連立方程式を解けば, $(\hat{\beta}_0, \hat{\beta}_1)$ が求まる.
ホワイトボードへ
その解は,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \hat{\beta}_0 = \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x}$$

で与えられる.

回帰直線

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \hat{\beta}_0 = \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x}$$

これらの値を, $y = \hat{\beta}_0 + \hat{\beta}_1 x$ に代入すると,

$$y = \frac{S_{xy}}{S_{xx}}(x - \bar{x}) + \bar{y}.$$

この式をyのxへの回帰直線と呼び, $\hat{\beta}_1$を回帰係数と呼ぶ.

例題 1(p.49)

(時間があれば...)