基礎セミナー
じっくり勉強すれば身につく統計入門

「じっくり勉強すれば身につく統計解析」を副題としたシリーズ全 3 巻がサイエンティスト社から刊行されました．タイトルは「医薬品開発のための統計解析，第 1 部 基礎，第 2 部 実験計画，第 3 部 非線形モデル」です．「じっくり勉強すれば身につく統計 ～ Excel，JMP による基礎から応用統計解析実務者コース」（SAS（株）JMP ジャパン事業部主催，年 12 回）のテキストとして使用されています．昨年の第 7 回定例会に先立って，この本をベースに「基本に戻ろうー基本統計量とデータの比較ー」を開催したところ好評であり．そこで，第 8 回定例会に先立って「回帰分析の基礎 ー95%信頼区間の活用ー」をテーマに開催することにいたしました．統計をじっくりと勉強して身に付けたいと思われる方々の参加をお待ちしています．

基本に戻ろう：回帰モデルとモデルの推定

橘田 久美子（スギ生物科学研究所）

日常的に何気なく使っている回帰分析を題材に，最小 2 乗法の復習をします．最小 2 乗法の本質を理解するために Excel のソルバーが適しています．Excel のソルバーを使いこなす技能を持つことにより，様々な統計解析の応用力を得ることができます．新しいことを学習するためには，簡単な問題から順次積み上げて行くことが効果的です．Excel シート上に入力されたデータを用いて，ソルバーを用いた回帰分析について実演します．同じデータを用いて Excel の LINEST 関数を使うことによっても回帰分析も簡単にできることも実演します．

基本に戻ろう：回帰直線モデル - 誤差を考慮した推定 -

杉本 典子

何回かの実験データから別々に回帰直線引くと微妙に異なる．回帰直線に関連する 95%信頼区間について理解を深めよう．このためには，Excel で実際の実験の結果をコンピュータ上で何回も繰り返しデータを発生させて回帰直線を Excel シート上で観察することが理解の助けになる．Excel を用いたコンピュータ上での“実験”シミュレーションで回帰直線の揺らぎを体験してみよう．回帰直線の 95%信頼区間，個々の観測データの 95%信頼区間が観測データ数によってどのように変化するのか，それらの 95%信頼区間の各種の応用を紹介する．ある反応 y_0 が得られたときの用量 x_{y0} の推定とその 95%信頼区間はどのように求めたらよいのか．Excel のゴールシークで求める方法を紹介する．

じっくり勉強すれば身につく
統計解析入門

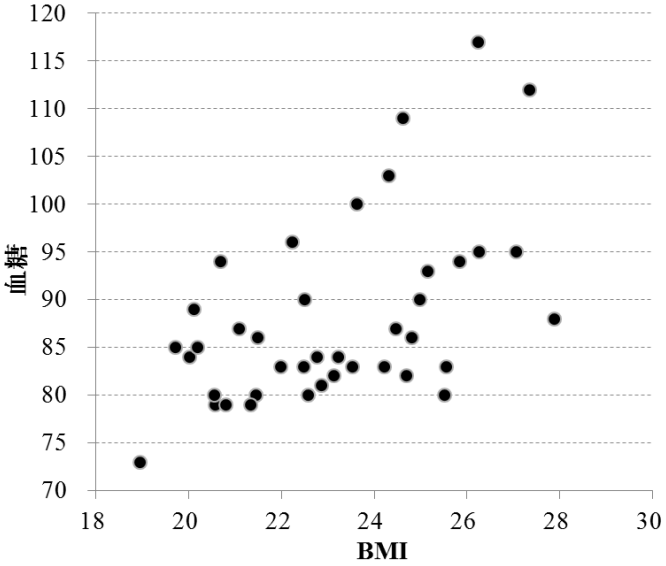
回帰モデルとモデルの推定

スギ生物科学研究所株式会社
橘田久美子

1. 回帰直線と回帰式

一般に肥満度が高くなると糖尿病に罹りやすくなるといわれる.
そこで，BMI (Body Mass Index, 体重[kg]÷身長[m]^2)と血糖の関係をグラフに表示してみた.

No.	BMI	血糖
1	26.3	95
2	27.9	88
3	26.2	117
4	25.6	83
5	20.1	89
6	20.7	94
	中略	
38	21.3	79
39	20.5	80
40	27.3	112
	BMI	血糖
平均	23.2	88.3
標準偏差	2.3	9.5

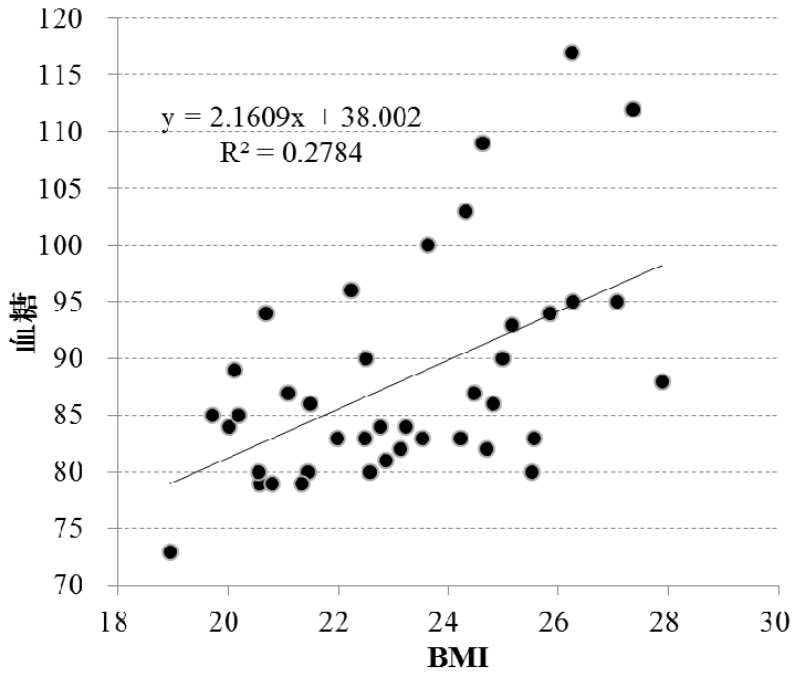


BMIが大きくなるほど血糖も高くなっていて，その関係は直線が当てはまりそう！

⇒相関係数
　回帰直線
を求めてみよう！

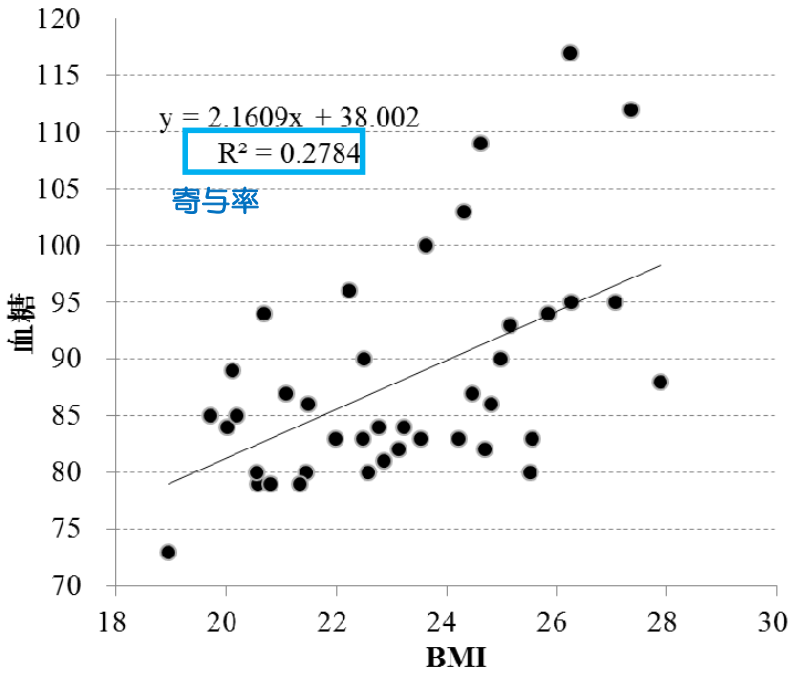
1. 回帰直線と回帰式

データ範囲を指定
↓
散布図を選択
↓
グラフを整形
↓
グラフ中のデータを指定
↓右クリック
[近似曲線の追加]から
・線形近似
・グラフに数式を表示する
・グラフにR-2値を表示する
を選択



3

1. 回帰直線と回帰式



Excelのグラフオプションでは相関係数 (r) ではなく寄与率 (R^2) が表示されるので注意！

相関係数を求めるときはR^2の値の平方根を計算する.

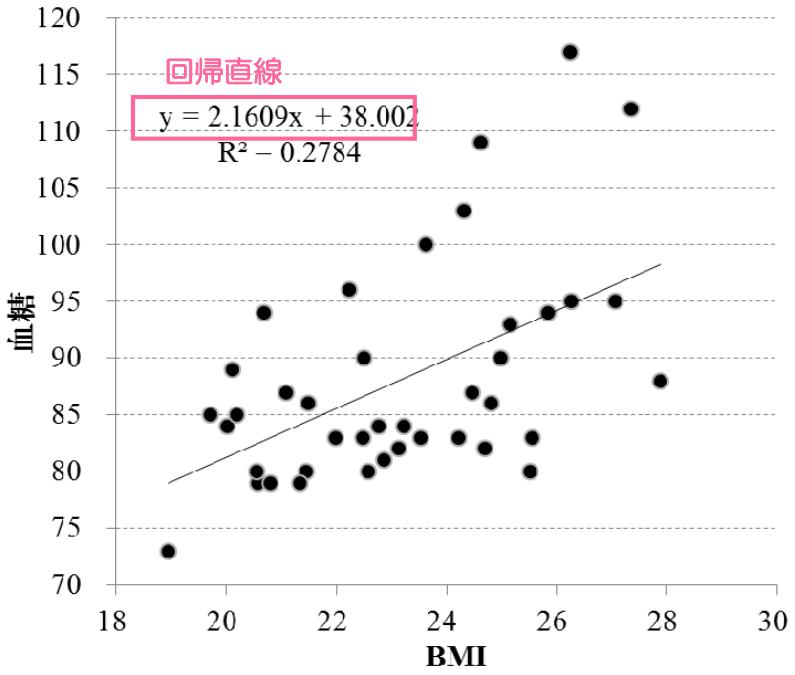
$\sqrt{0.2784} \approx 0.538$

BMI (x) と血糖 (y) の間には直線関係がありそう.

* 寄与率については後のスライドで説明します

4

1. 回帰直線と回帰式



回帰直線

y= 38.002 + 2.1609x

BMI (x) と血糖 (y) の間に直線関係があるならば，回帰直線を用いてBMI (x) から血糖 (y) を予測できる．

血糖 (=予測したい対象)
　目的変数，従属変数

BMI (=予測に使う変数)
　説明変数，独立変数

説明変数が1つ
　⇒**単回帰分析**

説明変数が複数個(2個以上)
　⇒**重回帰分析**

と呼ぶ．

2. 回帰モデル

BMIが x の人をたくさん集めて血糖値 y を調べた．

同じ x の人の血糖値 y は同じ値を示さない．これは x (BMI) が同じであっても個々の値には個人差 (誤差)が含まれるからである．

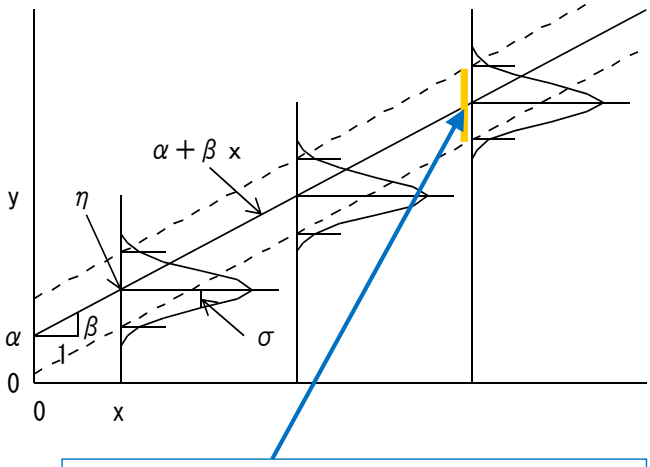
y はその母平均 η (イータ) を中心として変化する．

その変化を誤差 ε (イプシロン) とし，平均が 0，標準偏差が σ の正規分布に従うと想定すると

$$y = \eta + \varepsilon = \alpha + \beta x + \varepsilon$$

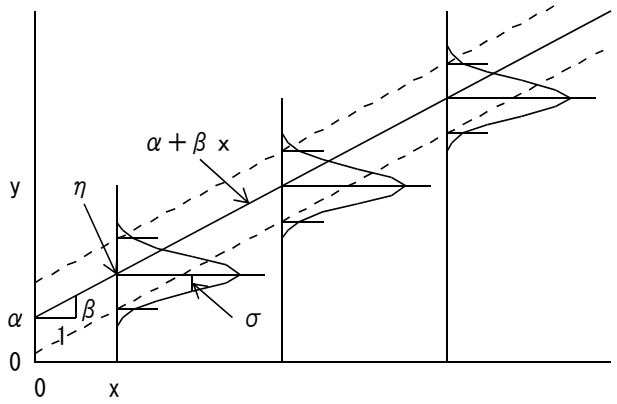
$$\varepsilon \sim N(0, \sigma^2)$$

で表される．



正規分布では $\eta \pm 1.96\sigma$ の間に95%が含まれる．
⇒y (=血糖値)の個々の観測値の95%はこの範囲に含まれる．

2. 回帰モデル



・y の方向には誤差があるが x に誤差はない.
・誤差は互いに独立で，各 y について相互に影響はない．(独立性)
・誤差の期待値は 0 である．(不偏性)
・誤差の大きさ σ は x の値にかかわらず一定である．(等分散性)
・誤差は正規分布にしたがう．(正規性)

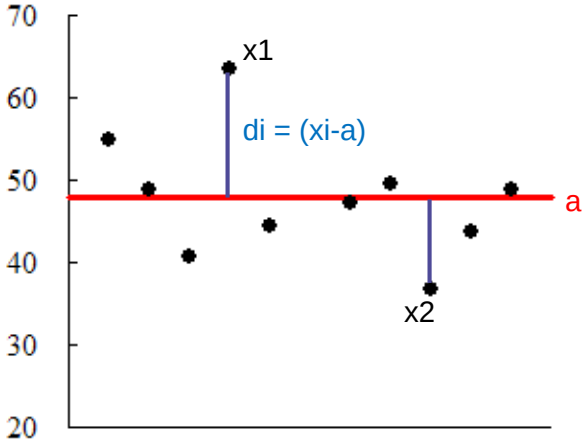
7

3. 最小2乗法による回帰式の推定

:復習:

観測値 xi が $N(\mu, \sigma^2)$ に従うとき，n 個の観測値から母平均 μ の推定値 a を求めたい．xi と a の距離 di (= xi-a) には +/- がある.

⇒ di を2乗すれば観測値の位置の影響 (平均値の上か下か)を排除できる.



$$S = \sum_{i=1}^{n} (xi-a)^2 \Rightarrow \min$$

となる a を求めればよく，この a が平均値となる.

この考え方が最小2乗法であった.

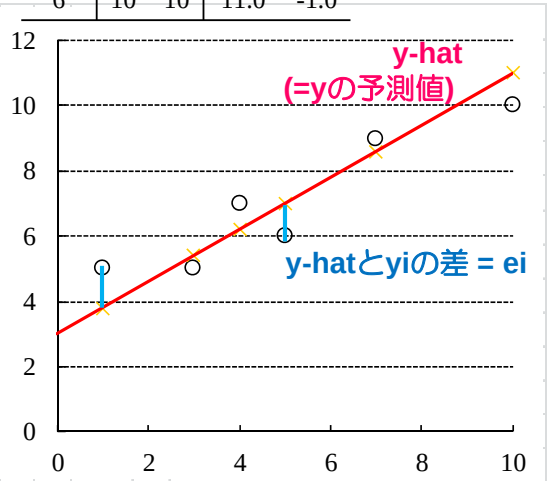
この最小2乗法という考え方を応用して
回帰式 $\alpha + \beta x$ の推定値 a，b
を求めてみる.

8

3. 最小2乗法による回帰式の推定

i	x	y	y-hat	e
1	1	5	2.9	2.1
2	3	5	4.7	0.3
3	4	7	5.6	1.4
4	5	6	6.5	-0.5
5	7	9	8.3	0.7
6	10	10	11.0	-1.0

a	2.0
b	0.9
S	8.2



回帰直線の一般式 y = a+ bx

直線式の各係数 a，bを変化させると係数にあわせてy-hat (yの予測値)が計算される.

ei は yi (観測値) と yi-hat (yの予測値)の差
ei = (yi – yi hat)

Sは各 ei を2乗した和
S = Σ (yi – yi hat)^2

として計算される.

実際に表のa，bを試行錯誤で変化させた場合のSとグラフの回帰直線の変化を観察してみよう.

S が最小になるように求められたa，bで作った式が回帰直線

この考え方が最小2乗法

9

4. ソルバーによる解法

最小2乗法の説明では，S を最小にするa，bを試行錯誤で求めた.

Excelのソルバーを使うと自動的にこの探索ができる.

ソルバーでは，指定したセル (目的セル，式が入力されている) の値を目標値にするような式の係数が自動的に計算される.

ソルバーを使用するためには，ソルバー機能がアドインされていることが必要.

[EXCEL2003]

トップメニュー [ツール]
↓
[アドイン]
↓
[有効なアドイン]でソルバーアドインを
チェック
↓
[OK]

[EXCEL2010]

トップメニュー [ファイル]
↓
[オプション]
↓
[アドイン]
↓
設定ボタンを押して [ソルバー]を設定

10

4. ソルバーによる解法

i	x	y	y-hat	e
1	1	5	2.9	2.1
2	3	5	4.7	0.3
3	4	7	5.6	1.4
4	5	6	6.5	-0.5
5	7	9	8.3	0.7
6	10	10	11.0	-1.0

a	2.0
b	0.9
S	8.2

:Excel2010の場合:

トップメニューの『データ』から[ソルバー]を選択

:Excel2003:の場合:

トップメニューの『ツール』から[ソルバー]を選択

目的セル

最適解を求めたいセル (= Sのセル) を選択

目標値

●最小 を選択

変化させるセル

a・b のセルを選択

↓

Excel2003では 実行
Excel2007では 解決
をクリック

↓

Sを最小にするようなa・b の値が自動的に計算され，解が得られる.

11

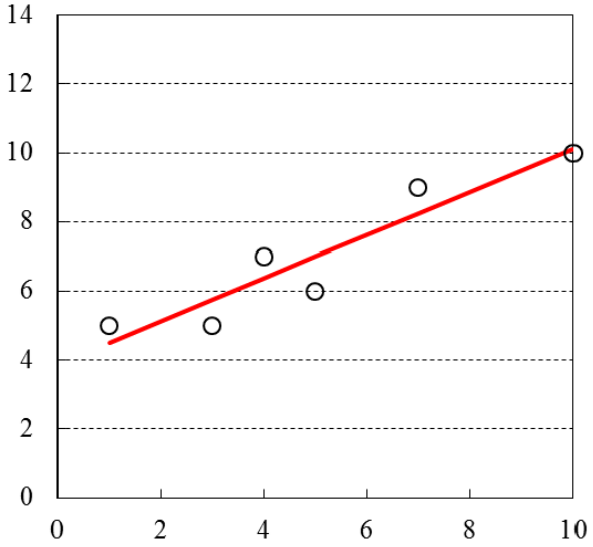
4. ソルバーによる解法

i	x	y	y-hat	e
1	1	5	4.52	0.48
2	3	5	5.76	-0.76
3	4	7	6.38	0.62
4	5	6	7.00	-1.00
5	7	9	8.24	0.76
6	10	10	10.10	-0.10
平均	5.00	**7.00**	**7.00**	**0.00**
平方和	50.00	**22.00**	**19.22**	**2.78**

相関係数	x	y	y-hat	e
x	1.000	0.935	1.000	**0.000**
y	0.935	1.000	0.935	0.355
y-hat	1.000	0.935	1.000	**0.000**
e	0.000	0.355	0.000	1.000

a	3.900
b	0.620
S	2.780

Sを最小にするようなa・b の値が自動的に計算され，解が得られた.



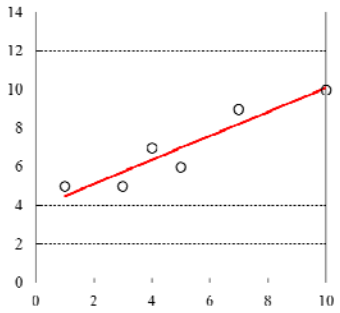
12

4. ソルバーによる解法

i	x	y	y-hat	e
1	1	5	4.52	0.48
2	3	5	5.76	-0.76
3	4	7	6.38	0.62
4	5	6	7.00	-1.00
5	7	9	8.24	0.76
6	10	10	10.10	-0.10
平均	5.00	**7.00**	**7.00**	**0.00**
平方和	50.00	**22.00**	**19.22**	**2.78**

相関係数	x	y	y-hat	e
x	1.000	0.935	1.000	**0.000**
y	0.935	1.000	0.935	0.355
y-hat	1.000	0.935	1.000	**0.000**
e	0.000	0.355	0.000	1.000

a	3.900
b	0.620
S	2.780



ソルバーで得られた回帰式から予測値 y-hat と残差を計算し，それらの関係を調べた．

- 実測値 y の平均と予測値 y-hat の平均はどちらも 7 で等しい．
- 残差 e の平均は 0
  ⇒ 残差の合計は 0
- 残差 e は説明変数 x，予測値 y-hat と無相関である．

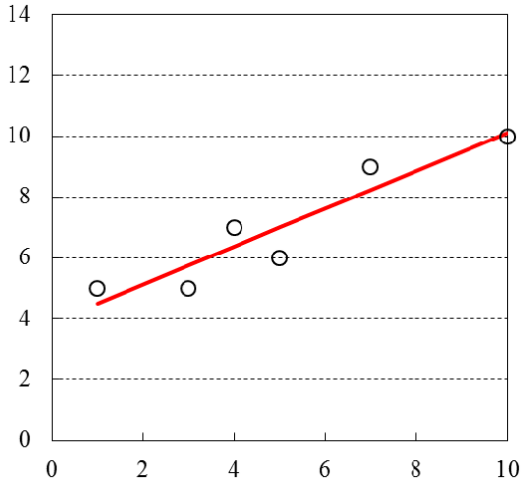
13

4. ソルバーによる解法

i	x	y	y-hat	e
1	1	5	4.52	0.48
2	3	5	5.76	-0.76
3	4	7	6.38	0.62
4	5	6	7.00	-1.00
5	7	9	8.24	0.76
6	10	10	10.10	-0.10
平均	5.00	**7.00**	**7.00**	**0.00**
平方和	50.00	**22.00**	**19.22**	**2.78**

相関係数	x	y	y-hat	e
x	1.000	0.935	1.000	**0.000**
y	0.935	1.000	0.935	0.355
y-hat	1.000	0.935	1.000	**0.000**
e	0.000	0.355	0.000	1.000

a	3.900
b	0.620
S	2.780

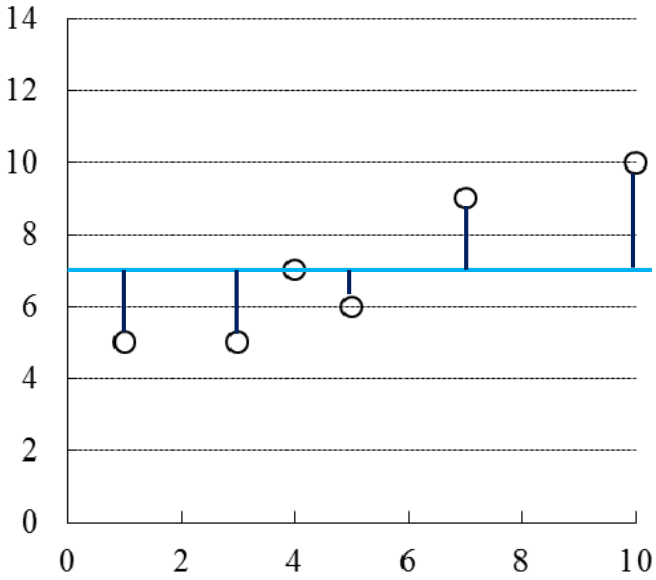


平方和 S = Σ(個々の値－平均値)^2

14

5. 平方和の分解とσ²の推定

i	x	y	y-hat	e
1	1	5	4.52	0.48
2	3	5	5.76	-0.76
3	4	7	6.38	0.62
4	5	6	7.00	-1.00
5	7	9	8.24	0.76
6	10	10	10.10	-0.10
平均	5.00	**7.00**	**7.00**	**0.00**
平方和	50.00	**22.00**	**19.22**	**2.78**



平方和 S = Σ(個々の値－平均値)²

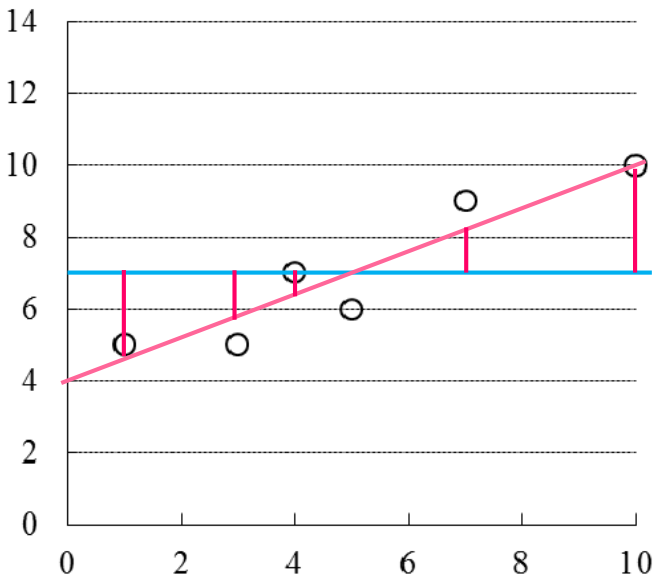
yの平方和 = S_T (総平方和)

⇒平均値と個々の値の差の平方和

15

5. 平方和の分解とσ²の推定

i	x	y	y-hat	e
1	1	5	4.52	0.48
2	3	5	5.76	-0.76
3	4	7	6.38	0.62
4	5	6	7.00	-1.00
5	7	9	8.24	0.76
6	10	10	10.10	-0.10
平均	5.00	**7.00**	**7.00**	**0.00**
平方和	50.00	**22.00**	**19.22**	**2.78**



平方和 S = Σ(個々の値－平均値)²

yの平方和 = S_T (総平方和)

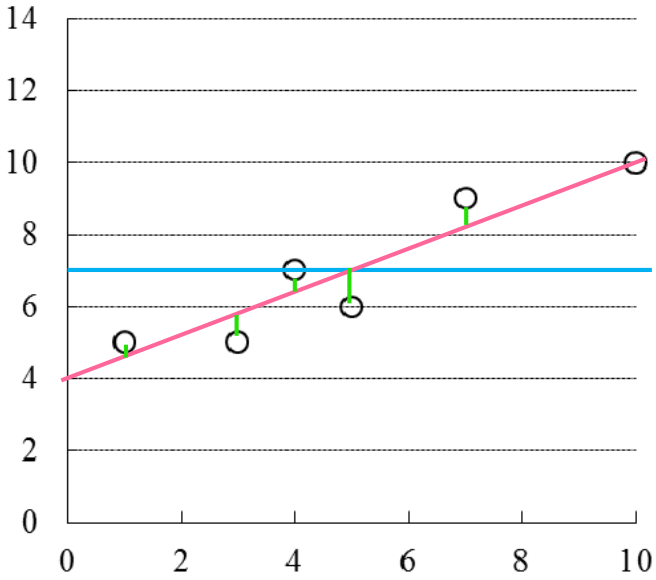
y-hatの平方和 = S_R (回帰平方和)

⇒平均値と回帰直線の差の平方和

16

5. 平方和の分解とσ²の推定

i	x	y	y-hat	e
1	1	5	4.52	0.48
2	3	5	5.76	-0.76
3	4	7	6.38	0.62
4	5	6	7.00	-1.00
5	7	9	8.24	0.76
6	10	10	10.10	-0.10
平均	5.00	**7.00**	**7.00**	**0.00**
平方和	50.00	**22.00**	**19.22**	**2.78**



平方和 S = Σ(個々の値－平均値)²

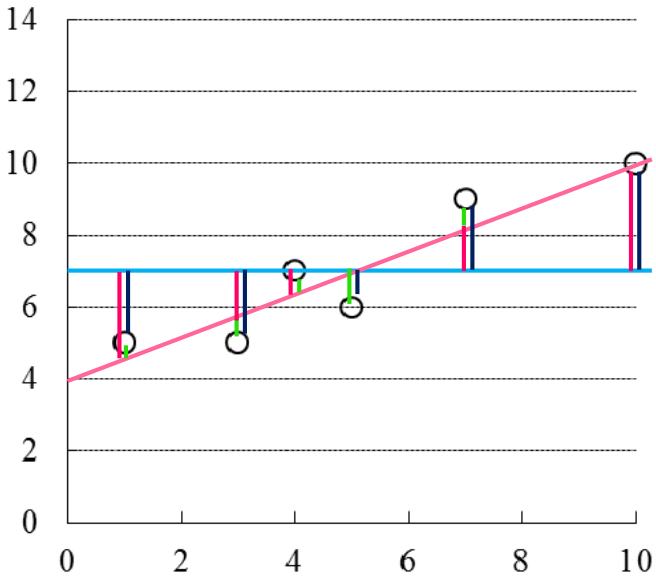
yの平方和　　= S_T (総平方和)
y-hatの平方和 = S_R (回帰平方和)
eの平方和　　= S_e (残差平方和)

⇒観測値と回帰直線の差の平方和

17

5. 平方和の分解とσ²の推定

i	x	y	y-hat	e
1	1	5	4.52	0.48
2	3	5	5.76	-0.76
3	4	7	6.38	0.62
4	5	6	7.00	-1.00
5	7	9	8.24	0.76
6	10	10	10.10	-0.10
平均	5.00	**7.00**	**7.00**	**0.00**
平方和	50.00	**22.00**	**19.22**	**2.78**



平方和 S = Σ(観測値－平均値)²

個々の yの値 yi とyの平均の差が
『 y の平均と回帰直線の差』
『回帰直線と個々の値の差』
に分解されることがわかる.

18

5. 平方和の分解とσ²の推定

i	x	y	y-hat	e
1	1	5	4.52	0.48
2	3	5	5.76	-0.76
3	4	7	6.38	0.62
4	5	6	7.00	-1.00
5	7	9	8.24	0.76
6	10	10	10.10	-0.10
平均	5.00	**7.00**	**7.00**	**0.00**
平方和	50.00	**22.00**	**19.22**	**2.78**

S_T = S_R + S_e

yの平方和 = S_T (総平方和)
y-hatの平方和 = S_R (回帰平方和)
eの平方和 = S_e (残差平方和)

$S_T = S_R + S_e$
22.00 = 19.22 + 2.78
の関係があることがわかる.

S_Rは y のばらつきのうち x の変化によって説明できる部分 (回帰直線を当てはめたことで説明できるようになった部分) の大きさを,
S_eは x の変化によって説明できない部分の大きさを，表している.

総平方和 S_T のうち，回帰平方和 S_R が占める割合を寄与率(または決定係数，R^2)という.

$R^2 = S_R / S_T = 19.22 / 22.00 = 0.874$

寄与率の平方根をとると相関係数と一致する.

19

5. 平方和の分解とσ²の推定

i	x	y	y-hat	e
1	1	5	4.52	0.48
2	3	5	5.76	-0.76
3	4	7	6.38	0.62
4	5	6	7.00	-1.00
5	7	9	8.24	0.76
6	10	10	10.10	-0.10
平均	5.00	**7.00**	**7.00**	**0.00**
平方和	50.00	**22.00**	**19.22**	**2.78**

相関係数	x	y	y-hat	e
x	1.000	0.935	1.000	**0.000**
y	0.935	1.000	0.935	0.355
y-hat	1.000	0.935	1.000	**0.000**
e	0.000	0.355	0.000	1.000

総平方和の自由度 ν_T
→ 6個のデータの平方和なので 6−1 = 5
回帰平方和の自由度 ν_R
→回帰を説明する変数が1つ (x) なので 1
残差平方和の自由度 ν_e
→『平均が0』『xと無相関』という2つの制約条件があるので 6−2 = 4

⇒ 5 = 1 + 4

自由度についても
$\nu_T = \nu_R + \nu_e$ と分解できる.

20

5. 平方和の分解とσ²の推定

i	x	y	y-hat	e
1	1	5	4.52	0.48
2	3	5	5.76	-0.76
3	4	7	6.38	0.62
4	5	6	7.00	-1.00
5	7	9	8.24	0.76
6	10	10	10.10	-0.10
平均	5.00	**7.00**	**7.00**	**0.00**
平方和	50.00	**22.00**	**19.22**	**2.78**

相関係数	x	y	y-hat	e
x	1.000	0.935	1.000	**0.000**
y	0.935	1.000	0.935	0.355
y-hat	1.000	0.935	1.000	**0.000**
e	0.000	0.355	0.000	1.000

回帰モデルでは誤差 ε のばらつきの大きさは標準偏差 σ で表すことができた.

誤差 ε は xでは説明できない部分なので, 残差 e で推定できると考えられる.

$S_e / \nu_e = V_e$ (残差平均平方 = 残差の分散) なので

$V_e = 2.78 / 4 =$ 0.695

↑誤差の分散 σ²の推定値

よって, σの推定値は

√0.695 = 0.834 ←残差標準偏差

21

5. a, b, S_Rなどの計算式

ここまで最小2乗法の考え方を使って回帰式の係数a・bを推定する考え方を数値を使って説明した. ここでは数式を少しだけ使用して最小2乗法について考えてみる.

回帰式の係数 a, bは,

残差 ei = yi – yi^hat

の2乗和 S が最小になるよう決められた.

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

そこで, S をa, b で偏微分して =0 とする連立方程式を解いてみる.

$$\frac{\partial S}{\partial a} = \sum_{i=1}^{n} 2e_i \frac{\partial e_i}{\partial a} = \sum_{i=1}^{n} 2e_i(-1) = -2\sum_{i=1}^{n} (y_i - (a + bx_i)) = 0$$

$$\frac{\partial S}{\partial b} = \sum_{i=1}^{n} 2e_i \frac{\partial e_i}{\partial b} = \sum_{i=1}^{n} 2e_i(-x_i) = -2\sum_{i=1}^{n} (y_i - (a + bx_i))x_i = 0$$

22

5. a, b, S_Rなどの計算式

この式を整理すると

$$na + \sum_{i=1}^{n} x_i b = \sum_{i=1}^{n} y_i \qquad \cdots 5.1式$$

$$\sum_{i=1}^{n} x_i a + \sum_{i=1}^{n} x_i^2 b = \sum_{i=1}^{n} x_i y_i \qquad \cdots 5.2式$$

が導かれる．これを正規方程式という．

5.1式の両辺を n で割ると

$$a + \frac{\sum x_i}{n} b = \frac{\sum y_i}{n}$$

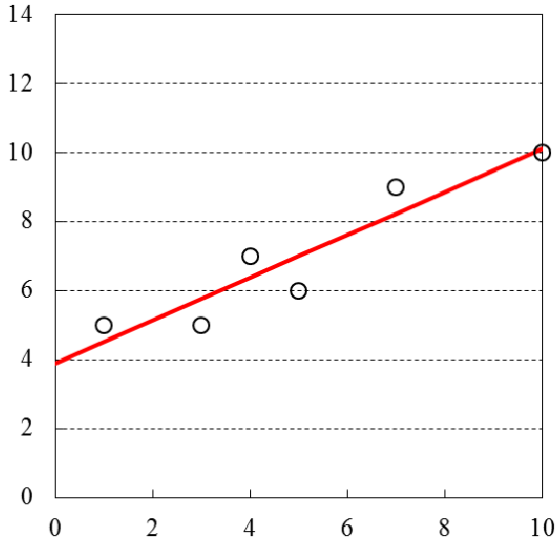
xの平均　yの平均　なので

$$a + \bar{x}.b = \bar{y}.$$

と書き換えられる．

aについて解けば

$a = \bar{y}. - b\bar{x}.$



23

5. a, b, S_Rなどの計算式

a は x=0 の縦軸と回帰直線の交点の y の値 (切片)を示している．

x が身長，y が体重を示す場合，aは身長が 0 の人の体重の平均を意味するが，このような値には意味がない．

このような場合は y = a + bx の a に

$a = \bar{y}. - b\bar{x}.$

を代入した

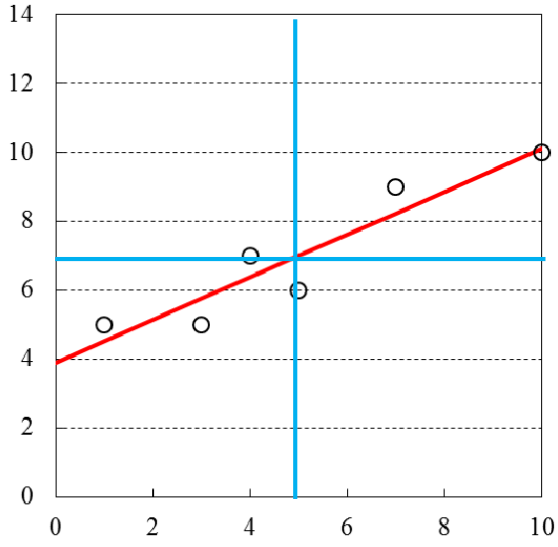
$y = a + bx = (\bar{y}. - b\bar{x}.) + bx$

$= \bar{y}. + b(x - \bar{x}.)$

という式を用いるのが良いであろう．

この式は，$x = \bar{x}.$をのとき $y = \bar{y}.$ となる．

つまり，回帰直線が x.と y. の交点，すなわち**重心 ($\bar{x}.$, $\bar{y}.$)を通る**ことを示している．



24

6. LINEST関数による解法

最小2乗法による回帰分析を手順を追って説明した.
今まで順に計算して求めた値は Excel の LINEST関数を使うと一度に計算できる.

i	x	y
1	1	5
2	3	5
3	4	7
4	5	6
5	7	9
6	10	10

	x	const	
係数	0.620	3.900	
se(係数)	0.118	0.681	
r^2	0.874	0.834	sd
F	27.655	4	fe
SR	19.220	2.780	Se

5行2列の空白セル (出力範囲) を選択する
↓
= LINEST (y範囲, x範囲, , TRUE)
を入力し, Ctrl・Shift キーを押しながら Enter をクリック
↓
解が得られる

25

まとめ

- 2つのデータの間の直線関係を知りたいときには回帰分析を行う.
- 直線の意味, 式の求め方, 寄与率 (R^2) や相関係数 (r) , 平方和, 残差標準偏差などの考え方をソルバーを使用して説明した.
- また, これらの値を一度に計算する方法としてLINEST関数を紹介した.
- これらの考え方は回帰分析の最も基礎となる考え方である.

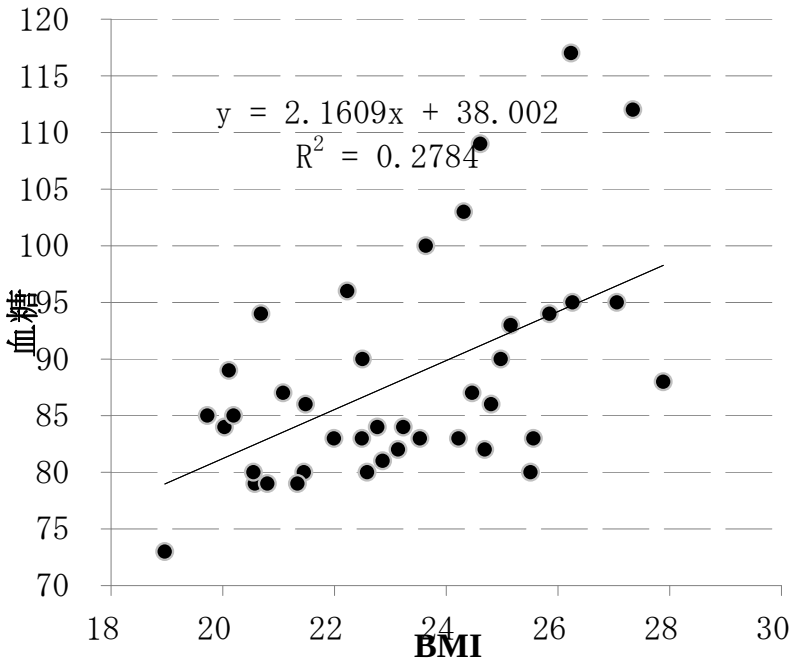
26

Excel シート 1

散布図と回帰式

No.	BMI	血糖
1	26.3	95
2	27.9	88
3	26.2	117
4	25.6	83
5	20.1	89
6	20.7	94
7	22.5	83
8	22.6	80
9	25.2	93
10	24.3	103
11	22.2	96
12	22.9	81
13	25.0	90
14	24.7	82
15	22.5	90
16	25.8	94
17	25.5	80
18	23.1	82
19	20.0	84
20	20.2	85
21	19.0	73
22	24.8	86
23	23.5	83
24	22.0	83
25	23.6	100
26	21.5	80
27	19.7	85
28	22.8	84
29	23.2	84
30	20.6	79
31	21.5	86
32	24.5	87
33	21.1	87
34	24.6	109
35	27.1	95
36	24.2	83
37	20.8	79
38	21.3	79
39	20.5	80
40	27.3	112

	BMI	血糖
平均	23.2	88.1
標準偏差	2.3	9.5
r	0.528	1.000

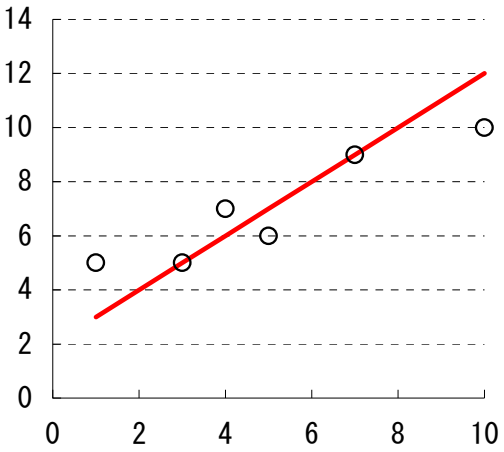


Excel シート 2

係数 a, bの値による回帰直線の変化

i	x	y	y^hat	e
1	1	5	3.0	2.0
2	3	5	5.0	0.0
3	4	7	6.0	1.0
4	5	6	7.0	-1.0
5	7	9	9.0	0.0
6	10	10	12.0	-2.0
平均	5	7		
平方和	50.0	22.0		
積和	31.0			

a	2.0
b	1.0
S	10.0



D4: =$H$4+$H$5*B4
E4: =C4-D4
H6: =SUMSQ(E4:E9)

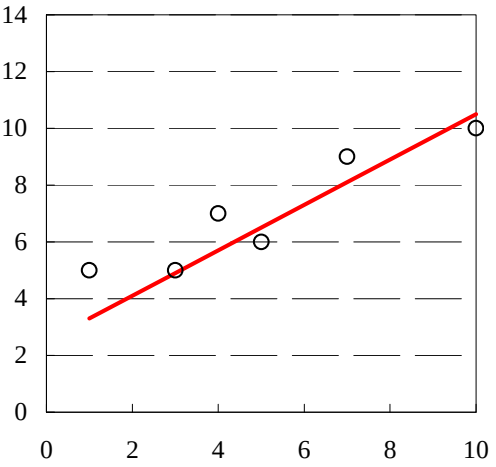
Excel シート 3

ソルバーによる解法

i	x	y	y-hat	e
1	1	5	3.30	1.70
2	3	5	4.90	0.10
3	4	7	5.70	1.30
4	5	6	6.50	-0.50
5	7	9	8.10	0.90
6	10	10	10.50	-0.50
平均	5.00	**7.00**	**6.50**	**0.50**
平方和	50.00	**22.00**	**32.00**	**4.40**

相関係数	x	y	y-hat	e
x	1.000	0.935	1.000	**-0.607**
y	0.935	1.000	0.935	-0.285
y-hat	1.000	0.935	1.000	**-0.607**
e	-0.607	-0.285	-0.607	1.000

a	2.500
b	0.800
S	5.900

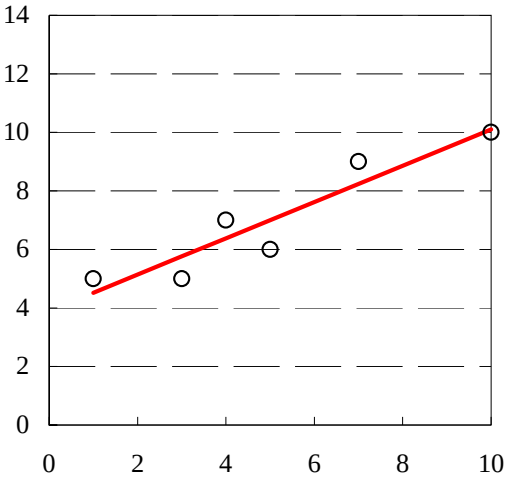


Excel シート 4

平方和の分解

i	x	y	y-hat	e
1	1	5	4.52	0.48
2	3	5	5.76	-0.76
3	4	7	6.38	0.62
4	5	6	7.00	-1.00
5	7	9	8.24	0.76
6	10	10	10.10	-0.10
平均	5.00	**7.00**	**7.00**	**0.00**
平方和	50.00	**22.00**	**19.22**	**2.78**

a	3.900
b	0.620
S	2.780



相関係数	x	y	y-hat	e
x	1.000	0.935	1.000	**0.000**
y	0.935	1.000	0.935	0.355
y-hat	1.000	0.935	1.000	**0.000**
e	0.000	0.355	0.000	1.000

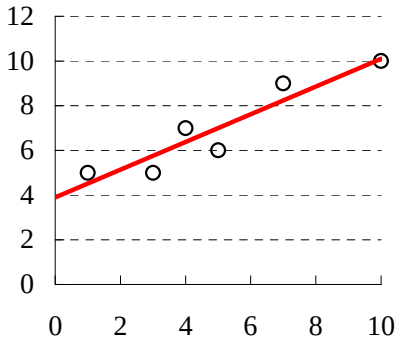
S_R/S_T= 0.874

相関係数の2乗= 0.8736

Excel シート 5

LINEST関数による計算

i	x	y
1	1	5
2	3	5
3	4	7
4	5	6
5	7	9
6	10	10
平均	5	7
平方和	50.0	22.0
積和	31.0	



LINEST関数の出力

	x	const	
係数	0.620	3.900	
se(係数)	0.118	0.681	
r^2	0.874	0.834	sd
F	27.655	4	fe
SR	19.220	2.780	Se

回帰直線モデル
誤差を考慮した推定

杉本　典子

1

はじめに

グリーン本

「じっくり勉強すれば身につく統計入門」がテーマ。
医薬の分野で統計的方法を適用する際の基本的な考え方を説明しており、入門者向け。
→本日は４章４節

ダウンロード先
http://www.scientist-press.com/12_278.html

2

(1) シミュレーション実験：実験内容

データには誤差が含まれるため、
観測値から得られたa, bの値にも誤差が含まれる.
これをシミュレーションでみてみよう

＜回帰直線モデルのシミュレーション＞

$$y = 20 + 1 \times x + e, \quad e \sim N(0, 15^2)$$

$$(\alpha = 20, \ \beta = 1, \ \sigma = 15)$$

上記の式に従うｙを発生させて
切片aと傾きｂの推定がどうなるか図でみる

(1) シミュレーション実験：Excel実験

X: 固定値
Y: (C8:C21) はE4:G4の母集団のパラメータに従い誤差を正規乱数として求めている.

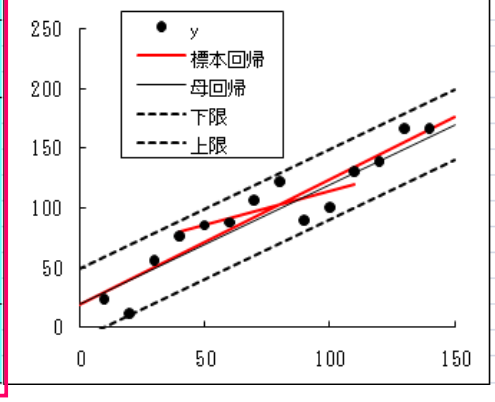
RAND：0～1の一様乱数を生成
NORMSINV：一様乱数を標準正規分布に従う乱数に変換

C8 のセルには

=$E$4+$F$4*B8+$G$4*NORMSINV(RAND())

= α + β *xi+ σ * N(0,1)

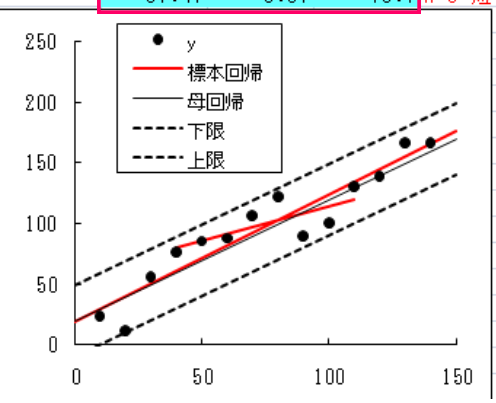
	A	B	C	D	E	F	G	H
1	表示4.4.1 回帰モデルのシミュレーション							
2								
3	x1	10			切片	傾き	σ	
4	dx	10		母回帰	20.00	1.00	15.0	
5				標本回帰	18.56	1.05	16.3	n=14 長い線
6					57.41	0.57	13.7	n=8 短い線
7			x	y				
8	1		10.0	23.6				
9	2		20.0	11.7				
10	3		30.0	56.1				
11	4		40.0	76.7				
12	5		50.0	85.4				
13	6		60.0	88.0				
14	7		70.0	106.6				
15	8		80.0	122.4				
16	9		90.0	89.9				
17	10		100.0	101.0				
18	11		110.0	130.8				
19	12		120.0	138.9				
20	13		130.0	166.4				
21	14		140.0	166.8				



(1) シミュレーション実験：Excel実験

			切片	傾き	σ	
x1	10	母回帰	20.00	1.00	15.0	
dx	10	標本回帰	18.56	1.05	16.3	n=14 長い線
			57.41	0.57	13.7	n=8 短い線

	x	y
1	10.0	23.6
2	20.0	11.7
3	30.0	56.1
4	40.0	76.7
5	50.0	85.4
6	60.0	88.0
7	70.0	106.6
8	80.0	122.4
9	90.0	89.9
10	100.0	101.0
11	110.0	130.8
12	120.0	138.9
13	130.0	166.4
14	140.0	166.8



観測値から求めた
回帰式係数a, b
誤差の標準偏差s
上：N=14
下：N=8(4～11)

・黒の直線　母回帰の直線
・赤の直線　n=14(長)とn=8(短)のデータから推定した回帰直線
・点線　母回帰直線の上下に1.96σの幅をつけたもの
　観測点の95%がこの範囲内に入ると考えられる線

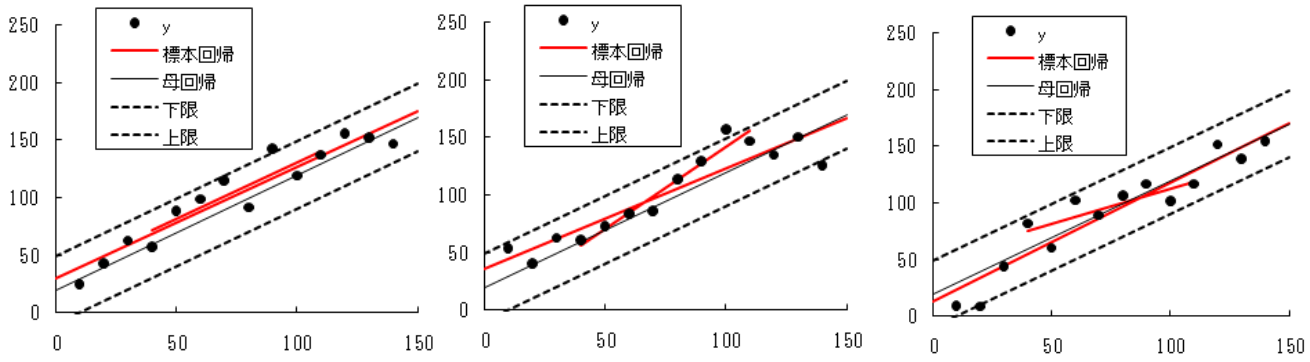
＜ＥＸＣＥＬで見る＞
F9を押す→ｙと回帰式のパラメータや散布図が変化

5

(1) シミュレーション実験：Excel実験

「F9」を押す→ｙと回帰式のパラメータや散布図が変化

実際のExcelで



観察してわかること
・推定値ａとｂ　は　ばらつくものだ
・ｎ＝14（長い線）とｎ＝8（短い線）の直線を
　比較　ｎ＝8がばらつきが大きいことがわかる

6

(2) a、 b の標準誤差

次に乱数を100 回発生させてa, b を求め,
その平均と標準偏差を求めるシミュレーション
もExcelでできます
（計算は省略：グリーン本とExcelを見て下さい）

		$n=14$		$n=8$	
		a	b	a	b
シミュレーション	平均	20.027	0.994	19.857	0.992
	標準誤差	8.523	0.091	18.107	0.214
理論値	平均	20.000	1.000	20.000	1.000
	標準誤差	8.468	0.099	18.151	0.231

シミュレーションを観察するｔｎが大きいほうが平均が理論値にちかいことがわかる
また、標準誤差はｎが大きい方が小さい

・標準誤差はどのように求めるのでしょうか

7

(2) a、 b の標準誤差

傾きbの標準誤差を求める式

傾きbの標準誤差s. e. [b]の理論値は次式で計算

$$s.e.[b] = \frac{\sigma}{\sqrt{S_{xx}}}$$

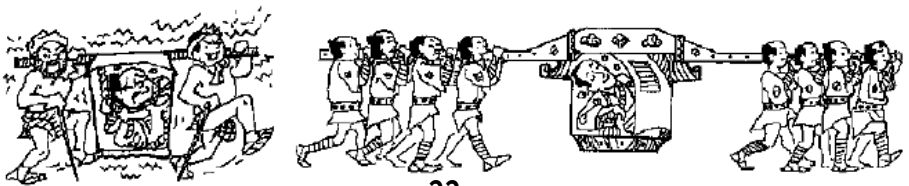
分母は

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x}.)^2$$

この式から
x_iの変化の範囲が広く、nが大きいほど
S_{xx}は大きくなる　→bの標準誤差は小さくなる

テキスト抜粋

式の誘導よりは式の意味を正しく理解することが大切である．そのためには，次の説明と挿絵が役立つであろう．
江戸時代の駕籠を考える．町人の乗る駕籠は 棒が短く，2 人で担ぐので，揺れが大きく，紐に しっかり しがみついて いなければならない．それに対して，大名行列の駕籠は棒が長く沢山の人で担がれるので，大名は うとうと しながら乗っている．



8

(2) a、 b の標準誤差

切片aの標準誤差を求める式

切片a の標準誤差の値は次式で求めることができる

$$s.e.[a] = \sqrt{\frac{1}{n} + \frac{\overline{x}.^2}{S_{xx}}} \cdot \sigma$$

計算例

		$n=14$		$n=8$	
		a	b	a	b
シミュレーション	平均	20.027	0.994	19.857	0.992
	標準誤差	8.523	0.091	18.107	0.214
理論値	平均	20.000	1.000	20.000	1.000
	標準誤差	8.468	0.099	18.151	0.231

N=14, 8 の場合はS_{xx} = 22750, 4200 なので

$s.e.[b] = \frac{\sigma}{\sqrt{S_{xx}}}$

$$s.e.[b] = \frac{15}{\sqrt{22750}} = 0.099, \qquad \frac{15}{\sqrt{4200}} = 0.231$$

切片a の標準誤差の値は次式で求めることができる

$$s.e.[a] = \sqrt{\frac{1}{14} + \frac{75^2}{22750}} \times 15 = 8.468, \qquad \sqrt{\frac{1}{8} + \frac{75^2}{4200}} \times 15 = 18.151$$

9

(3) β の仮説検定と区間推定

傾きβの仮説検定

回帰直線の傾きを表わすβの推定値と
標準誤差が求められた
次は 帰無仮説 H_0: β = 0 の検定 を考える

↓

xとyの間の回帰直線が水平かどうかを検定するもの

↓

母相関係数が0かどうかの検定と同等

＜検定方法＞

bがその標準誤差s. e. [b] の何倍であるか

$$t = \frac{b}{s.e.[b]}$$

10

(3) βの仮説検定と区間推定

傾きβの仮説検定：例

以下のデータで検定してみると

	L	M	N	O	P
3	i	x	y	y-hat	e
4	1	1	5	4.52	0.48
5	2	3	5	5.76	-0.76
6	3	4	7	6.38	0.62
7	4	5	6	7.00	-1.00
8	5	7	9	8.24	0.76
9	6	10	10	10.10	-0.10
10	平均	5.00	**7.00**	**7.00**	**0.00**
11	平方和	50.00	**22.00**	**19.22**	**2.78**

	L	M	N	O
21		x	const	
22	係数	0.620	3.900	
23	se(係数)	0.118	0.681	
24	r^2	0.874	0.834	sd
25	F	27.655	4	fe
26	SR	19.220	2.780	Se

=DEVSQ(M4:M9)

$$s.e.[b] = \frac{\sigma}{\sqrt{S_{xx}}} \qquad s.e.[b] = \frac{0.834}{\sqrt{50.0}} = 0.118$$

$$t = \frac{b}{s.e.[b]} \qquad t = \frac{0.620}{0.118} = 5.259$$

σは分からないので，
LINEST関数で求めた
推定値s = 0.834を用いる
*LINEST関数を利用すると
aとbの標準誤差も計算し
てくれている

自由度が4のt分布でp値=0.006
→これから回帰直線は水平ではない
すなわちyはxによって変化することが分かる

11

(3) βの仮説検定と区間推定

傾きβの95% 信頼区間

傾きβの95% 信頼区間は

$$b - t(0.05, \nu_e)s.e.[b] < \beta < b + t(0.05, \nu_e)s.e.[b]$$

$$0.620 - 2.776 \times 0.118 < \beta < 0.620 + 2.776 \times 0.118$$

$$0.293 < \beta < 0.947$$

＊回帰直線の切片を表わすα についても
同様の検定と区間推定をすることができる

12

(4) 予測値とyの区間推定

$\eta = \alpha + \beta x$の区間推定

回帰モデル $y = \eta + \varepsilon = \alpha + \beta x + \varepsilon, \qquad \varepsilon \sim \mathrm{N}(0,\ \sigma^2)$

パラメータ α, β の推定値a，bとその標準誤差も求められた
次は $\eta = \alpha + \beta x$ の区間推定をする

$$\hat{\eta} = \hat{y} = a + bx = \overline{y}. + b(x - \overline{x}.)$$

から $\hat{\eta}$ の分散を考える→分散V[$\hat{\eta}$] は分散の加法性により

$$\begin{aligned} \mathrm{V}[\hat{\eta}] &= \mathrm{V}[\overline{y}.] + (x - \overline{x}.)^2 \mathrm{V}[b] \\ &= \frac{1}{n}\sigma^2 + (x - \overline{x}.)^2 \frac{1}{S_{xx}}\sigma^2 \\ &= \left(\frac{1}{n} + \frac{(x - \overline{x}.)^2}{S_{xx}}\right)\sigma^2 \end{aligned}$$

η の95% 信頼区間推定は σ^2 をVe で置き換えて

$$\eta \sim a + bx \pm t(0.05, \nu_e)\sqrt{\left(\frac{1}{n} + \frac{(x - \overline{x}.)^2}{S_{xx}}\right) V_e}$$

13

(4) 予測値とyの区間推定

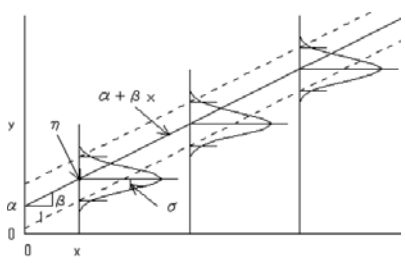
η の95% 信頼区間の性質

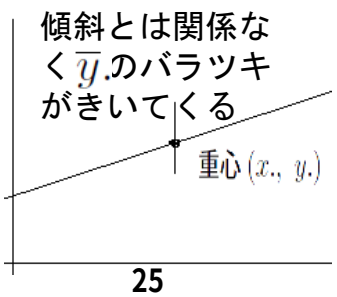
$$\eta \sim a + bx \pm t(0.05, \nu_e)\sqrt{\left(\frac{1}{n} + \frac{(x - \overline{x}.)^2}{S_{xx}}\right) V_e}$$

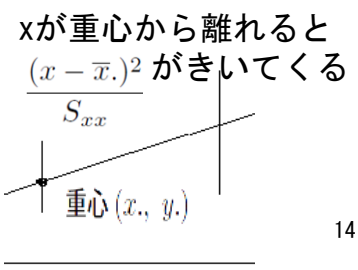
①n が大きくなると区間は狭くなる（1/n→小）

②xの範囲が広くなると区間は狭くなる（1/Sxx→小）

③平均値に近いxの区間は狭くなる（$(x - \overline{x}.)^2$ →小）





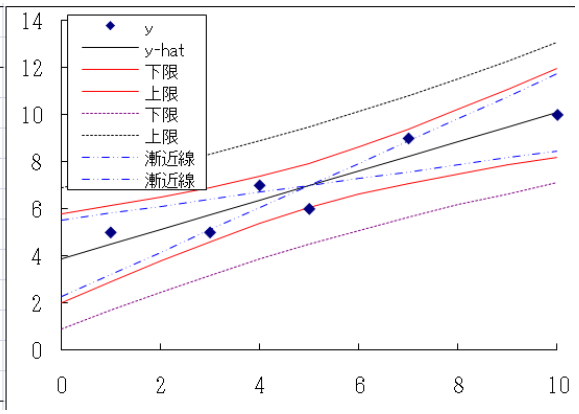


14

(4) 予測値とyの区間推定

$\eta = \alpha + \beta x$の区間推定：例 x=0～10

			η の信頼限界		y の信頼限界	
x	y	y-hat	下限	上限	下限	上限
0		3.90	2.01	5.79	0.91	6.89
1	5	4.52	2.91	6.13	1.70	7.34
2		5.14	3.78	6.50	2.45	7.83
3	5	5.76	4.61	6.91	3.18	8.34
4	7	6.38	5.38	7.38	3.86	8.90
5	6	7.00	6.06	7.94	4.50	9.50
6		7.62	6.62	8.62	5.10	10.14
7	9	8.24	7.09	9.39	5.66	10.82
8		8.86	7.50	10.22	6.17	11.55
9		9.48	7.87	11.09	6.66	12.30
10	10	10.10	8.21	11.99	7.11	13.09



内側の赤の双曲線
母回帰直線 $\eta = \alpha + \beta x$ が含まれる確率が95%となる範囲を表わす。
漸近線

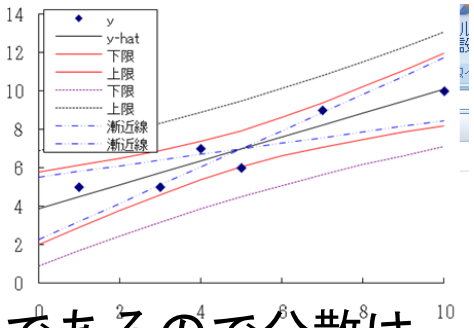
$$y = \bar{y}. + b_L(x - \bar{x}.)$$
$$y = \bar{y}. + b_U(x - \bar{x}.)$$

bL： bの下側信頼限界値0.293
bU： bの上側信頼限界値0.947

η の信頼区間の幅はx = x. で最小→xがx. から離れるにつれて広くなる.
データとしてとったxの範囲の外側で、xに対する η を推定する（回帰直線を外挿する）のは注意が必要

(4) 予測値とyの区間推定

観測値yの信頼区間



右図の外側の双曲線の曲線が
観測値 y の信頼区間の例
個々の観測値の95%信頼区間

観測値yは η に ε が加わったものであるので分散は

$$\begin{aligned} \mathrm{V}[y] &= \mathrm{V}[\hat{\eta}] + \mathrm{V}[\varepsilon] \\ &= \left(\frac{1}{n} + \frac{(x - \bar{x}.)^2}{S_{xx}}\right)\sigma^2 + \sigma^2 \\ &= \left(1 + \frac{1}{n} + \frac{(x - \bar{x}.)^2}{S_{xx}}\right)\sigma^2 \end{aligned}$$

yの区間推定は

$$y \sim (a + bx) \pm t(0.05, \nu_e)\sqrt{\left(1 + \frac{1}{n} + \frac{(x - \bar{x}.)^2}{S_{xx}}\right)V_e}$$

$$\eta \sim a + bx \pm t(0.05, \nu_e)\sqrt{\left(\frac{1}{n} + \frac{(x - \bar{x}.)^2}{S_{xx}}\right)V_e}$$

n が大きくなると， y=a+bx±1.96sの直線に近づく

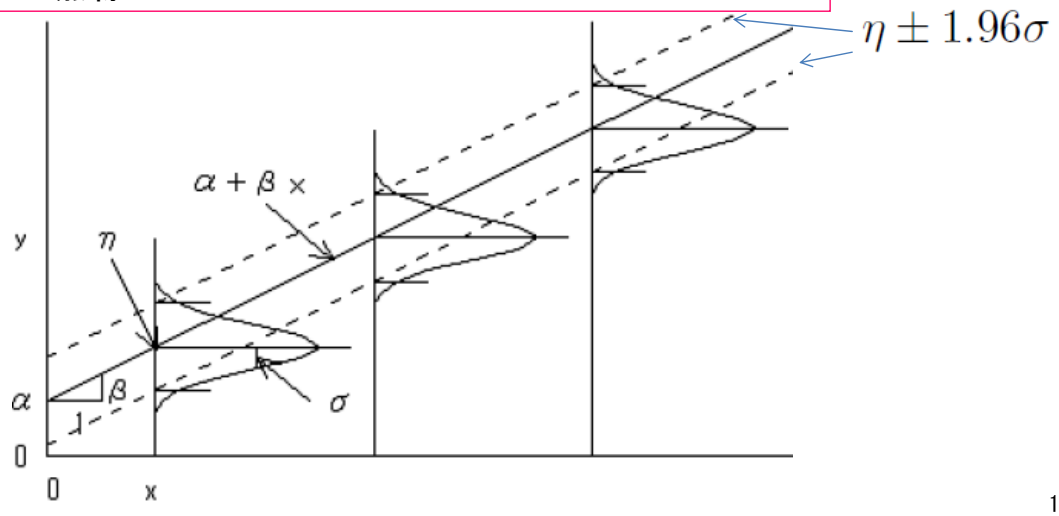
(4) 予測値とyの区間推定

観測値yの信頼区間

yの区間推定

$$y \sim (a+bx) \pm t(0.05, \nu_e)\sqrt{\left(1+\frac{1}{n}+\frac{(x-\bar{x}.)^2}{S_{xx}}\right)V_e}$$

n が大きくなると，y=a+bx±1.96sの直線に近づく
とはこの点線



17

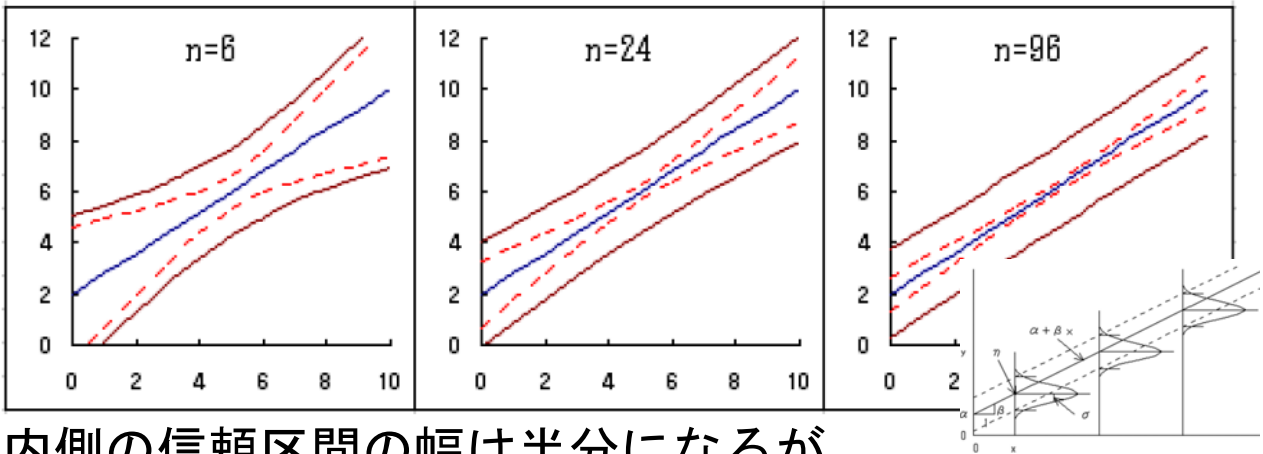
(4) 予測値とyの区間推定

ηとyの信頼区間の違い

■ η の信頼区間：
ある説明変数xにおける母平均 η =a+bxの信頼区間（平均値の信頼精度）

■y の信頼区間：
ある説明変数xにおける測定値yの信頼区間（測定値が得られる範囲）

n により2つの双曲線がどう変化するか



内側の信頼区間の幅は半分になるが，
外側の信頼区間の幅の変化は少ないことが分かる

18

(5) 逆推定

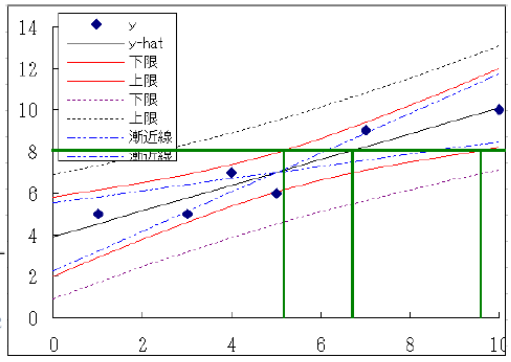
yの期待値 η が8になるxを推定したい
→例：y=8の水平線と回帰直線の交点のxを読み取る

計算で求めるには回帰式 $y = a + bx = 3.9 + 0.62x$
を変形してyに8を代入

$$x = \frac{y-a}{b} = \frac{8-3.9}{0.62} = \frac{4.1}{0.62} = 6.61$$

この推定を逆推定という.

■xの信頼区間が知りたいとき
同様にy=8の水平線と内側の
双曲線の交点を読み取れば良い



↑幅が違う

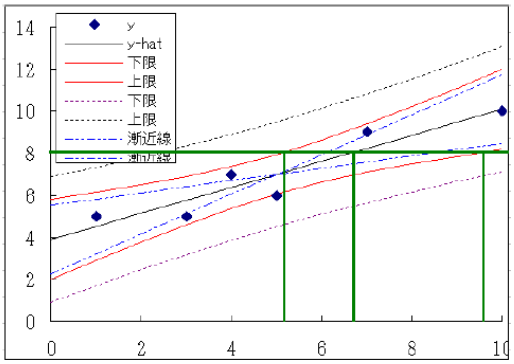
$$\eta \sim a + bx \pm t(0.05, \nu_e)\sqrt{\left(\frac{1}{n} + \frac{(x-\overline{x}.)^2}{S_{xx}}\right)V_e}$$

を解けば良いが，xの2次式で面倒な計算が必要となる
→ソフトの力を借りる（ゴールシーク）

19

(5) 逆推定

■内側の曲線の逆推定の区間推定の使われ方



→つまり η の信頼区間
→使われ方
横軸のxが薬剤の投与量で，縦軸のyが薬効のとき

複数の被験者に投与量を変化させた薬効を
プロットしたとする.
平均の薬効が8になる投与量はどの位かを求めた
ことになる.

20

(5) 逆推定

■外側の曲線の逆推定の区間推定の使われ方
→ｙの信頼区間

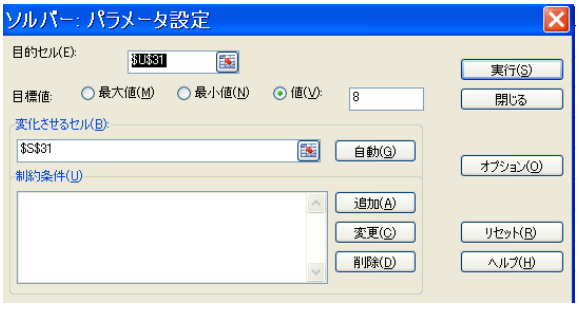
→化学分析の検量線を求めるため回帰直線を
あてはめるような場合

濃度xが既知の検体について分析し、
分析計の表示値yを求めて回帰分析しその信頼区間のグラフを描く.
濃度が未知の検体の表示値yから検体の濃度xを予測したい.
測定値yが得られた検体の集まりの濃度を推定するのではなく，今測定した検体の濃度を推定したい.
測定値yには誤差が含まれており，誤差を含む信頼区間(外側の曲線)を使い区間推定をする必要がある.

(5) 逆推定

ゴールシークによる逆推定

ソルバー： Excel2007［データ］>［ソルバー］の中
目的セル：U31
目標値：8
変化させるセル：S31



			ｍの信頼限界		ｙの信頼限界	
x	y	y-hat	下限	上限	下限	上限
6.61		8.00	6.92	9.08	5.44	10.56

→Excelでデモ

・ｘ＝6.61、 95%信頼区間（ 5.09 ， 9.38 ）
→点推定値と上下信頼区間の差
6.61 − 5.09 = 1.52, 9.38 − 6.61 = 2.77
となり等しくない.

まとめ

回帰式y = a + bx の推定方法
↓
データには誤差が含まれるため,
得られた推定値a, bの値にも誤差が含まれる.
—これをシミュレーションで認識
— a, bの標準誤差
— yの推定値の標準誤差
—逆推定（y からx を推定したい場合）