回帰分析とソルバー

- 2 つの変量を散布図に描いた場合、2 変量の間に関係が深いと点の散らばりが狭い範囲に集中する
- 狭い範囲に散らばったデータ点を特定の関数で表現して縮約することを回帰とよぶ
- 線形関数への回帰は、結果を人間が理解しやすいため比較的よく使われる
- 関数への回帰を求める方法は一つではないが、Excel 自体は最小 2 乗法だけをサポートしている
- ソルバーと呼ばれる Excel の機能を使うと、その他の様々な回帰を行うことができる
- といっても、”Graphs are essential to good statistical analysis.”(Anscombe, F. J., 1973)であることには変わりない。絵を描かずに計算するような愚行はしないように
-

1. 散布図と回帰直線

表 1　2変量の測定例

X1	4	5	6	7	8	9	10	11	12	13	14
Y1	4.26	5.68	7.24	4.82	6.95	8.81	8.04	8.33	10.84	7.58	9.96

表 1 は、2 つの変量 X と Y の計測結果である。この計測結果から 2 つの変量の間にどのような関係があるかを考えてみよう。まず、いつものとおり散布図を描いてみると図 1 のようになる。全体としてなんとなく右上がりの絵になっており、X と Y の間にはなんとなく関係があるような感じである。

ここまでは特に関数を特定化する必要がないのだが、回帰という考え方は X と Y が特定の関数に従っていると仮定して適切なパラメータを決めることなので、何らかの定式化が必要になる。もちろんどのような関数を選ぶかは分析者に任されているわけなので、猿のように常に線形関数を選ぶ必然性は全くないことに注意。

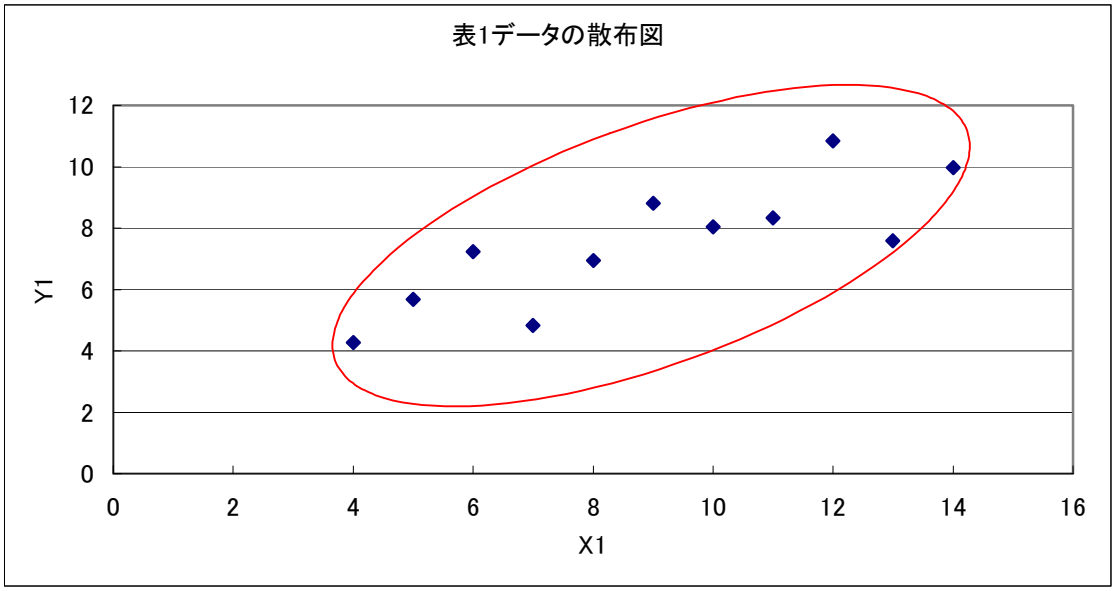


図 1　表 1 データの散布図

このケースでは図 1 を見る限り、なんとなく右上がりの直線的関係になっていそうな感じであるので、とりあえず

$$Y=\alpha+\beta X \qquad (1)$$

という線形的な関係を仮定してみることにする。このように関数を特定化すると、残された課題はいかに切片 α と傾き β をデータに合わせて決めるかだけである。もしデータがすべて(1)式を満たすなら（＝直線上にならんでいるなら）話は簡単なのだが、実際には図 1 に示したようにデータは直線上に並んでいないのが普通である。つまり、実際のデータに対して(1)式を考えると、一般的には i 番目のデータの組(X_i, Y_i)に対しては

$$Y_i = \alpha + \beta X_i + \varepsilon_i \qquad (2)$$

のように表すことができる。直線 $Y=\alpha+\beta X$ 上に乗っているデータに関しては $\varepsilon=0$ となるが、外れているデータでは $\varepsilon \neq 0$ である。ε が大きいほど直線とデータとの乖離は大きくなることがわかる。個別の ε_i を最小化するには特定のデータ点を通る直線を引けばそれで終わりであるが、沢山データ点がある場合には 1 つの点だけを最適化してもダメである。どのように α と β を定めると一番「点の分布に近い」直線となるかという問題に関する解は実は一つではないが、よく使われる方法に最小２乗法と呼ばれる方法がある。

この方法が最適な推定値を与えるためにはかなりいろいろな前提条件があるのだが、それらの前提条件は現実にそういうデータが多いからではなく一番簡単に推定を行えるという

理由で設定されていること[1]に注意が必要である。

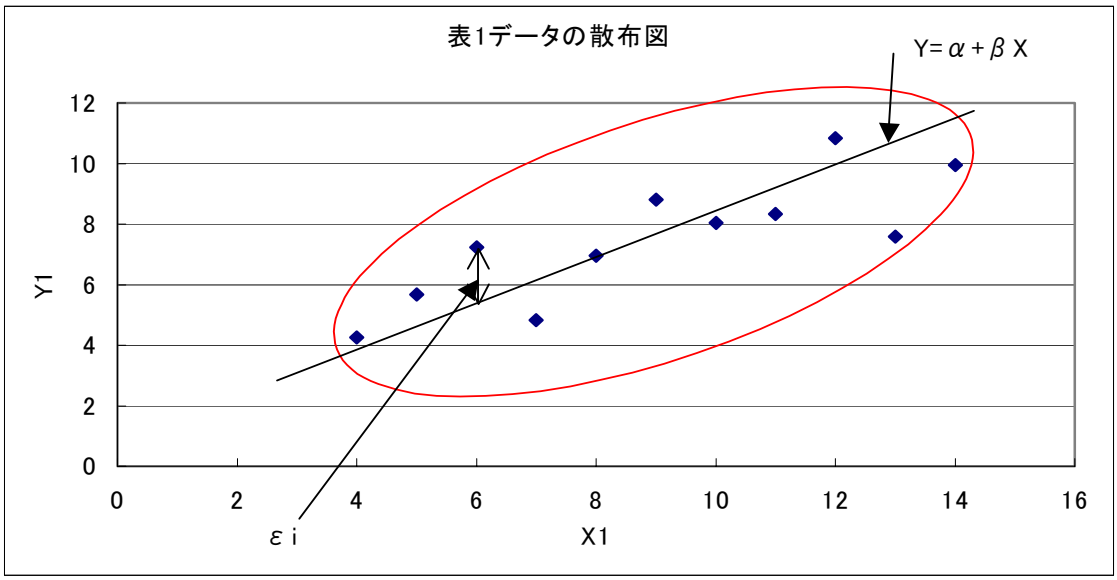


図 2　ε の図示

2. 最小２乗法

最小２乗法では、n 個の観測値があるときに

$$\sum_{i=1}^{n}\varepsilon_i = \sum_{i=1}^{n}(Y_i - \alpha - \beta X_i)^2 \qquad (3)$$

を最小化するように α と β の値を定める。２乗和を最小化するので「最小２乗法」というわけである。この方法によって適切な α と β が決まるためには、ε についてかなり強い仮定が必要になる。その仮定とは ε を確率変数として考えた場合に

$E(\varepsilon_i) = 0$　　平均は 0 で一定

$V(\varepsilon_i) = \sigma^2$　　分散は σ ２で一定

$E(\varepsilon_i \varepsilon_j) = 0$　　for i≠j　ε には系列相関がない

[1] Amemiya, Takeshi,　*Advanced Econometrics*, 1985 の 2 ページには"not because we believe that they are satisfied in most applications, but because they make a convenient starting point." と書いてある。そんなものだ。

の 3 つである。当然のことながら、実際のデータでこのような条件がうまく成立することはほとんどない。

2.1 Excel の機能をもちいて最小２乗法推定を行う

Excel には最小２乗法によってパラメータ推定を行う複数の方法がある。2 変量の場合には、散布図に「近似曲線の追加」という機能で回帰直線と回帰方程式を記入することが可能である。この方法は、実際のデータと求めた回帰直線がダイレクトに比較できる上、非線形の関数に対する回帰も行えるので 2 変量の場合は試してみる値打ちがある。もう一つの方法は、分析ツールに入っている「回帰分析」を使う方法である。こちらは説明変数の個数が増えても扱えるところが便利であるが、所詮おまけのツールなので専用の統計パッケージのように沢山の情報を出力することはできない。

2.1.1 近似曲線の追加

散布図中のデータ系列を選択して右ボタンでメニューを表示すると、図 3のような「近似曲線の追加」という項目がある。これを選択すると、**エラー! 参照元が見つかりません。**のダイアログが表示され、６種類の関数を使った近似曲線をグラフに追加することができる。推定した数式をグラフ内に表示するためには、[オプション]タブを開いて指定すればよい。

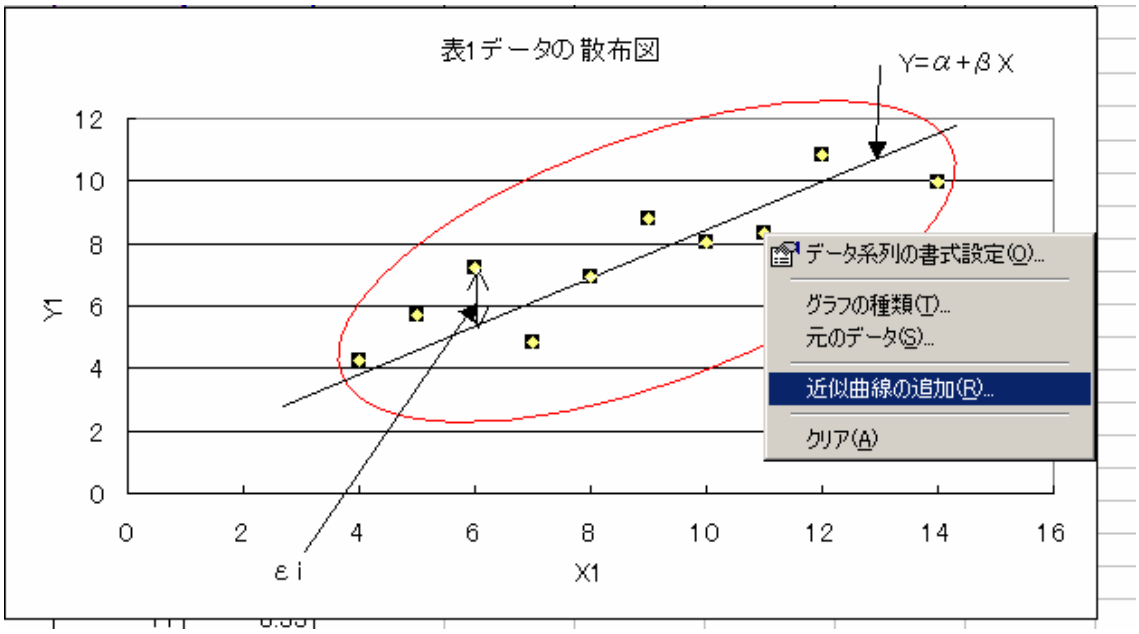


図 3　近似曲線の追加メニュー

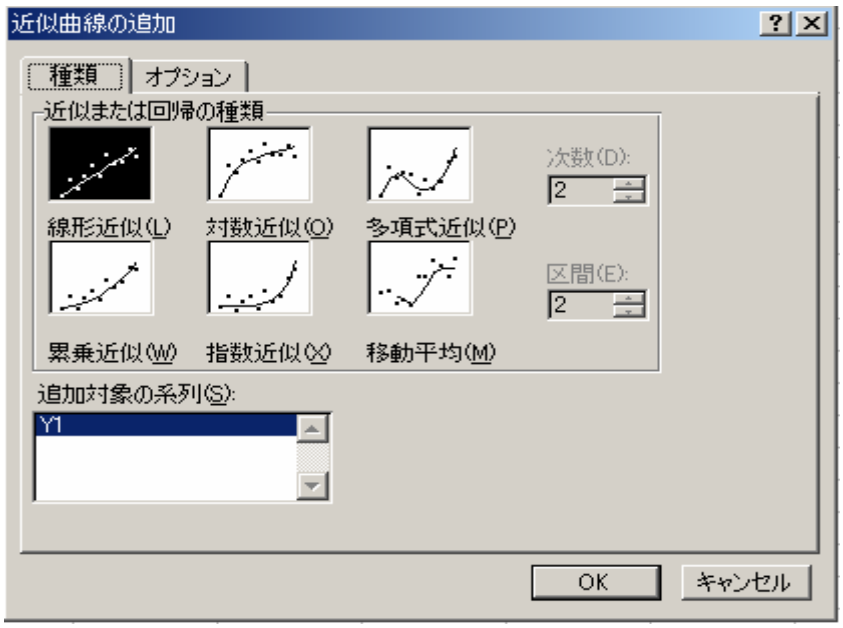


図 4　近似曲線の追加

この方法で近似曲線を追加すると図 5のようになる。左上に表示されている数式は、追加した近似曲線の（この場合は直線だが）式を表している。この式より、α=3, β=0.5 であることがわかる。表 1 のデータは線形関数で近似したためこのような式となっているが、近似曲線の追加では非線形の関数も選ぶことができるので、実際のデータと見比べながらいろいろ試してみることができる。

また、オプションとしてデータを外挿して推定式のグラフを描くこともできるので非線形の関数あてはめを行うときは試してみるとよい。サンプルがある区間でだけよい近似となっているが、ちょっとデータが外れるととんでもなく暴れるような fitting を行っているのに気づかず将来予測などに使う危険を減らすことができる。

2.1.2 分析ツール（回帰分析）

ヒストグラムと同様に、分析ツールを使って回帰分析を行うことができる。この場合は重回帰（説明変数の数が増えた場合）も推定できるが、使えるモデルは線形モデルだけになる。回帰分析ツールの設定画面は、図 6のような画面である。

入力 Y 範囲:　　従属変数の値が入っている列を指定する

入力X範囲：　　独立変数の値が入っている列（複数列可）を指定する

だけで基本的な推定は行える。表 1 のデータで推計を行った結果を図 7に示す。近似曲線の追加よりかなりいろいろな情報が出力されていることが分かる。特に t 値や棄却確率が計

算されているのは推計結果の評価時にはうれしい物である。

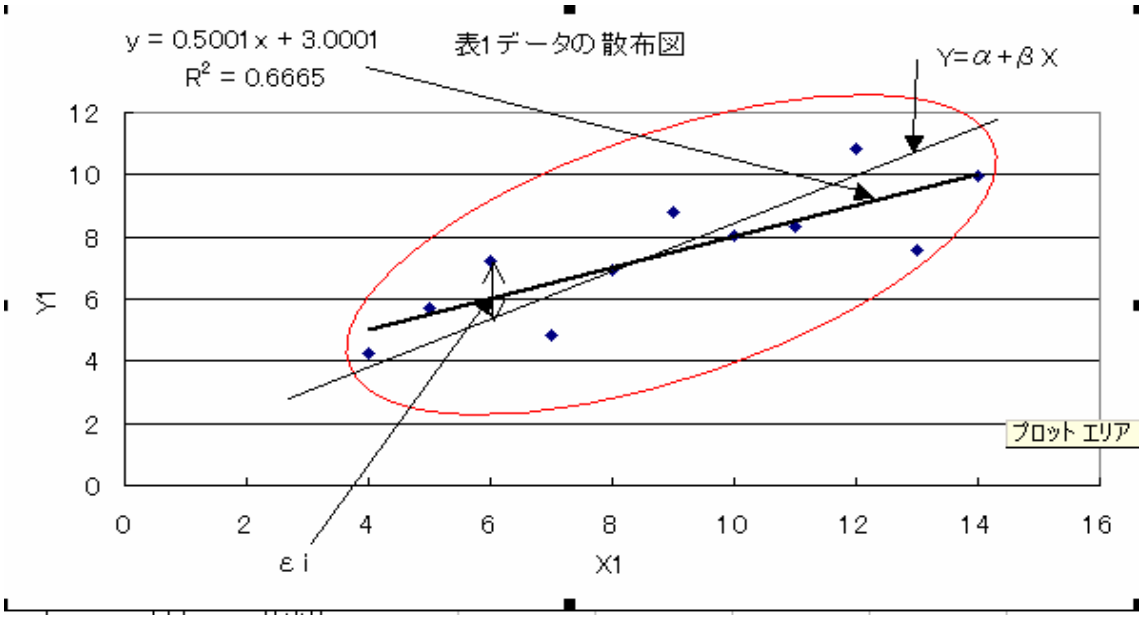


図 5 近似曲線を追加し、回帰式を表示した

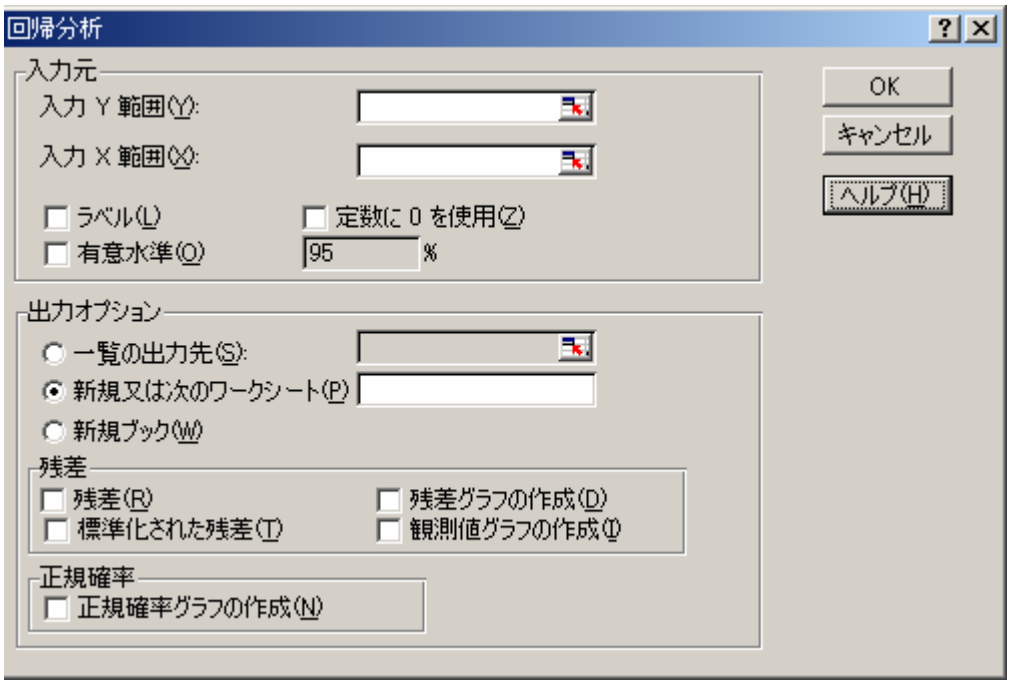


図 6　回帰分析ツールの設定画面

	A	B	C	D	E	F	G	H	I
1	概要								
2									
3	回帰統計								
4	重相関 R	0.816421							
5	重決定 R2	0.666542							
6	補正 R2	0.629492							
7	標準誤差	1.236603							
8	観測数	11							
9									
10	分散分析表								
11		自由度	変動	分散	測された分散	有意 F			
12	回帰	1	27.51	27.51	17.98994	0.00217			
13	残差	9	13.76269	1.529188					
14	合計	10	41.27269						
15									
16		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
17	切片	3.000091	1.124747	2.667348	0.025734	0.455735	5.544447	0.455735	5.544447
18	X 値 1	0.500091	0.117906	4.241455	0.00217	0.23337	0.766812	0.23337	0.766812
19									

図 7　回帰分析ツールでの出力

練習問題 1

anscombe.xls には(Anscombe, 1973)から引用した 4 系統の測定データ（測定 1 から 4）が含まれている。この 4 系統のうち、測定 1 と測定 2 のデータについて以下の作業を行いなさい

(1)「回帰分析ツール」を用いて $Y=\alpha+\beta X$　という関数に対する回帰分析を行いなさい

(2)散布図をまず描いて、「近似曲線の追加」機能を用いて近似曲線を追加しなさい。ただし、関数形は散布図を見て自分で決めなさい

(3)散布図を描かずに線形関数を決めうちして回帰分析を行うことが正しいかどうか考察しなさい

2.2 ソルバーを使って最小 2 乗法推定を行う

Excel 組み込みの機能を使わずに最小 2 乗法で回帰直線を決定するためには、(3)式を最小化するような α と β を自分で定める必要がある。もちろん統計学のテキストを読めばわかるように、上のような単純な最小 2 乗法については解析的に最適な α と β を求めることが可能であるが、ここでは Excel の持っているソルバーという機能を使って(3)を最小化してみることにする。ソルバーでは、制約条件つきの最大化、最小化、特定値への収束演算が行えるため(3)を最小化することもできるわけである。

2.2.1 ソルバーの使い方

ソルバーを使うためには、分析ツールの時と同様に[ツール]-[アドイン]からソルバーを選んで組み込む必要がある。組み込みが終わると、[ツール]メニューの中に「ソルバー」という選択肢が現れる。ソルバーを起動した時に表示されるダイアログは図 8である。各部分の内容はおおまかにいうと

- 目的セル: 最適化の対象となる計算式の入ったセルを指定する。ここで指定したセルの値が、「目標値」で指定した値になるまで繰り返し演算を行う
- 目標値:　　目的セルをどのような値にするかを指定する。最大値、最小値、指定値への収束が選べる。目的セルとして指定できるのは数式が入力されている単一のセルである
- 変化させるセル:　　目的セルに入っている計算式に影響を与えるセルで、目標値を達成するために操作したいデータの入っているセルを指定する。最小２乗法の場合は、係数が入っているセルを指定することになる。一つ以上のセルを指定することができるが、当然のことながらここで指定したセルの値を変えると目的セルの値も変わるようなセルを指定すること
- 制約条件: 非負条件や整数条件などを指定できる

となる。

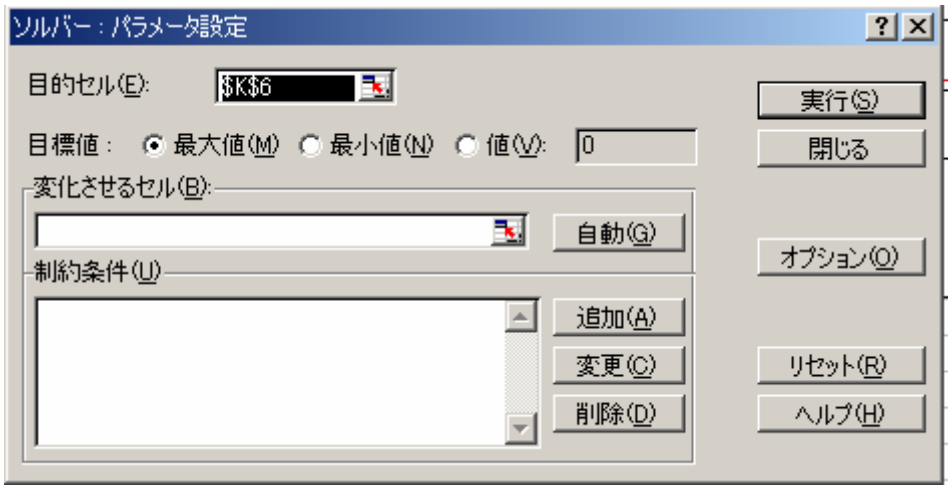


図 8　ソルバーのパラメータ入力画面

例題 1. 表 1 のデータとソルバーをつかって Y= α + β X という回帰方程式の α と β を推定してみる。

Step1 αとβの初期値を入れるセルを決める。図 9 の例では、αの値が入っているセルは **B14**, βの値が入っているセルは **B15**、初期値はどちらも 1 である。この 2 つのセルがソルバーでの「変化させるセル」になる。

Step2 αとβの値および X, Y の値を使って各データについてε^2の値を計算するセルを作る（図 10）

Step3 $\Sigma \varepsilon^2$を計算する計算式を **C13** に入れる。これがソルバーでの「目的セル」になる（図 11）。

Step4 ソルバーにStep 1、3で用意したセルを指定して、目標値を「最小値」にする（図 12）。

Step5 **「実行」ボタンを押すとソルバーが最適化を行い、結果をワークシートに記入する。αとβの値がそれぞれ 3 と 0.5 になっている（**

図 13）。

	A	B	C
1	X	Y	
2	4	4.26	
3	5	5.68	
4	6	7.24	
5	7	4.82	
6	8	6.95	
7	9	8.81	
8	10	8.04	
9	11	8.33	
10	12	10.84	
11	13	7.58	
12	14	9.96	
13			
14	α	1	
15	β	1	
16			

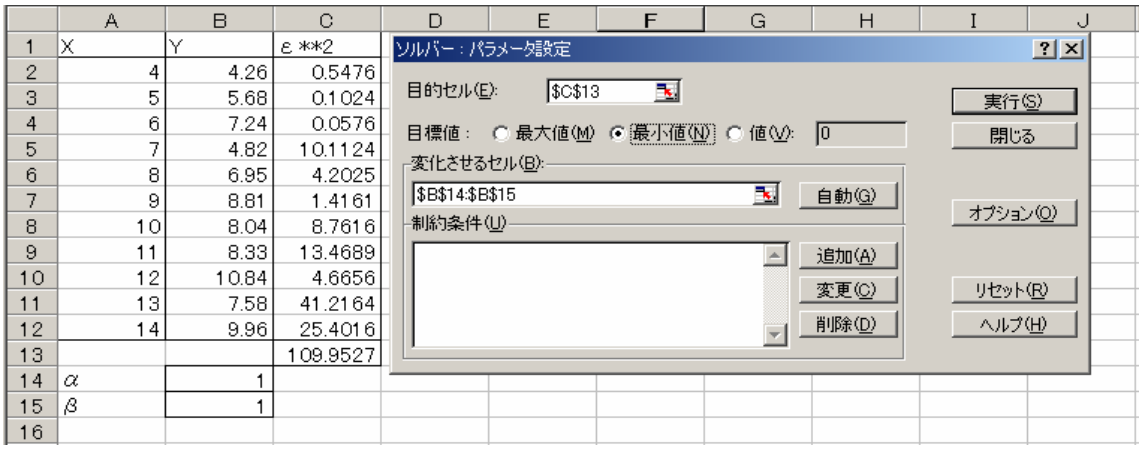
図 9 αとβの初期値を入力する

	A	B	C	D	E
1	X	Y	ε **2		
2	4	4.26	=(B2-B14-B15*A2)^2		
3	5	5.68	0.1024		
4	6	7.24	0.0576		
5	7	4.82	10.1124		
6	8	6.95	4.2025		
7	9	8.81	1.4161		
8	10	8.04	8.7616		
9	11	8.33	13.4689		
10	12	10.84	4.6656		
11	13	7.58	41.2164		
12	14	9.96	25.4016		
13					
14	α	1			
15	β	1			
16					

図 10　α と β の値を使って ε ² を計算する

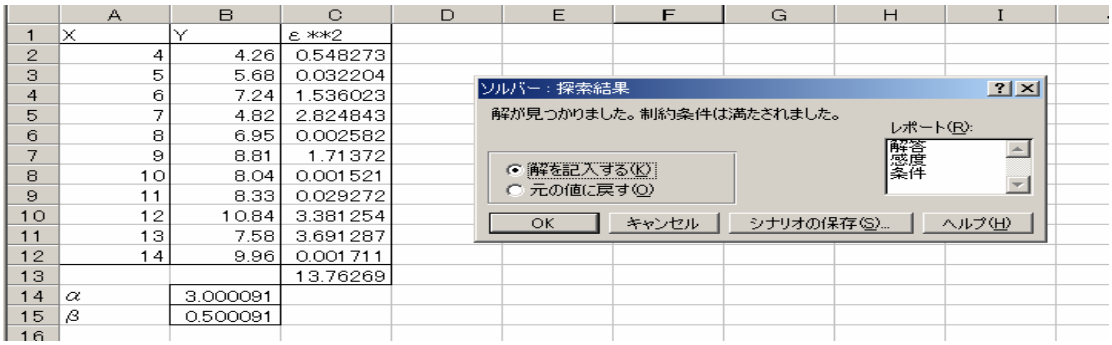
	A	B	C	D
1	X	Y	ε **2	
2	4	4.26	0.5476	
3	5	5.68	0.1024	
4	6	7.24	0.0576	
5	7	4.82	10.1124	
6	8	6.95	4.2025	
7	9	8.81	1.4161	
8	10	8.04	8.7616	
9	11	8.33	13.4689	
10	12	10.84	4.6656	
11	13	7.58	41.2164	
12	14	9.96	25.4016	
13			=SUM(C2:C12)	
14	α	1		
15	β	1		
16				

図 11　Σ ε 2 を計算するセルを作る。これが目的セル



	A	B	C
1	X	Y	ε **2
2	4	4.26	0.5476
3	5	5.68	0.1024
4	6	7.24	0.0576
5	7	4.82	10.1124
6	8	6.95	4.2025
7	9	8.81	1.4161
8	10	8.04	8.7616
9	11	8.33	13.4689
10	12	10.84	4.6656
11	13	7.58	41.2164
12	14	9.96	25.4016
13			109.9527
14	α	1	
15	β	1	
16			

図 12　ソルバーに目的セル・変化させるセル・目標値を設定



	A	B	C
1	X	Y	ε **2
2	4	4.26	0.548273
3	5	5.68	0.032204
4	6	7.24	1.536023
5	7	4.82	2.824843
6	8	6.95	0.002582
7	9	8.81	1.71372
8	10	8.04	0.001521
9	11	8.33	0.029272
10	12	10.84	3.381254
11	13	7.58	3.691287
12	14	9.96	0.001711
13			13.76269
14	α	3.000091	
15	β	0.500091	
16			

図 13ソルバーによる最適解。 α=3, β=0.5

2.2.2 重み付き最小２乗法

公表されているクロスセクションデータは大部分が個票ではなく度数分布表の形でまとめられている。また、個票であったとしてもサンプルウェイトはサンプルごとに異なることは少なくない。このようなケースでは度数分布の階級毎あるいはサンプルごとに重みを変えて回帰分析を行う必要がある。この場合は、最小２乗法の最小化問題を(3)からサンプルごとの重みを付けたものに変えればよい。つまり、サンプル i のウェイト（度数分布の場合は度数）を w_i とすると、

$$\sum_{i=1}^{n} w_i \varepsilon_i^{\,2} = \sum_{i=1}^{n} w_i \left(Y_i - \alpha - \beta X_i\right)^2 \qquad (4)$$

を最小化するような α と β を決定する問題に帰着する。Excel の組み込み分析ツールでは、このような重み付きの最小２乗法を使った回帰分析はできないが、ソルバーを使えばこのような分析も容易に行える。

3. 最小２乗法がうまく行かないケース

まず、anscombe.xls の測定 3 および測定 4 のデータを用いて散布図を描き線形の近似曲線を追加してみた結果を図 14、図 15に示す。これらの図からわかることは、最小２乗法は前提条件を満たさないデータに対しては非常に不安定な結果を返す危険性があるということである。どちらのデータ例も ε の分散が一定という条件を全く満たしていないが、このような異常値があったときに最小２乗法では適切な推定結果を得られないことが広く知られている。図をよく見ると、どの場合も α=3、β=0.5 となっていることが分かるであろう。つまり測定１から４はすべて線形回帰を行うと同じ係数となる数値例となっている。散布図を描くと全く適合していないにもかかわらずこのような結果がでてしまうのは、最小２乗法では ε を 2 乗した上で扱っているからである。２乗という演算は遠くのものをより遠くに、すぐ近くのものはより近くに写す効果がある。そのため平均から外れた異常値が大きな効果をもってしまい、全体に直線が引っ張られて散布図から大きくはずれた回帰直線となって現れている。

このような不安定性は最小２乗法の(3)式から出てくる結果であり、(3)式を最小化の目的関数として使っている限り解決しない。この問題に対してロバスト推定[2]と呼ばれる方法がいくつか開発されている。Excel の分析ツールなどではサポートしていないが、これもソルバーを使うことによって容易に実験[3]してみることが可能である。

3.1 最小絶対値法[4]

(3)では 2 乗して加えていたが、そもそも何故単純に加算するのではいけないのだろうか？これは ε が正負両方の値を取りうるため、単純に加算すると最小化問題が解を持たないからである（データ点から上に離れればどんどん $\Sigma\ \varepsilon$ は小さくなる）。2 乗しているのは、単に全てのデータを非負の数にしておくための数学的便法にすぎない。

上で見たように 2 乗という演算が遠くのものをより遠くに離す性質を持っているため異常値の影響が大きくなっていたのであるから、負の値を取らずになおかつ遠くのものを必要以上に遠くに追いやらない演算を使えばよりいい結果が得られる可能性がある。ここではそのような演算として絶対値を取ることを考えてみる。(3)と同じ記号を用いて書くと、

$$\sum_{i=1}^{n}|\varepsilon_i| = \sum_{i=1}^{n}|y_i - \alpha - \beta x_i| \qquad (5)$$

---

[2] いろんな人がいろんな関数で試してみている。詳しくは前掲の(Amemiya 1985)などを参照のこと。この授業では、(3)の代わりに別の関数を用いることによって違う結果が得られるということを扱うにとどめる。

[3] ロバスト（robust）推定は、単純な最小２乗法とは違いその性質が複雑である。Excel でもソルバーを使うことによってそのサワリ程度を試すことはできるが、真面目に使う場合は専用の統計パッケージを使う方がやはり楽である。

[4] 最小絶対値法　という名前が一般的にあるわけではない。最小２乗法と対比してここで仮につけた名前

となる。この(5)を最小化するようにαとβを決めると推定結果はどのように変わるであろうか。Anscombe の数値例１と３を使って計算した結果を図 16、図 17に示す。性質のいいデータでは２つの推計方法の結果はほとんど一致するが、性質の悪いデータでは最小絶対値法の方がなんとなくもっともらしい推定結果となっている。

より一般的なロバスト推定の方法としては、M-Estimator と呼ばれる方法がある。これは、最小２乗法の(3)を別の関数に置き換えるものであり、

$$\sum_{i=1}^{n} \rho(\varepsilon_i) \qquad (6)$$

を最小化するように係数を定める。ρは適切な性質を持った関数である。しかし、これについてはこの授業の範囲を超えるのでこれ以上扱わない。あとは計量経済学の授業を履修すること。

練習問題 2

anscombe.xls の「練習問題 2」シートに入っているデータは、サンプルごとにウェイトがことなっている（サンプルごとに異なるサンプリング率でサンプル調査されたと考えればよい)。ウェイトを考慮して線形回帰を行った場合と、ウェイトを考慮しなかった場合の両方について回帰係数を求めなさい。またこのことから、サンプル調査でたまたま抽出率が低く（＝復元ウェイトが大きく）なっていたサンプルが異常値を取った場合の影響を考察しなさい。

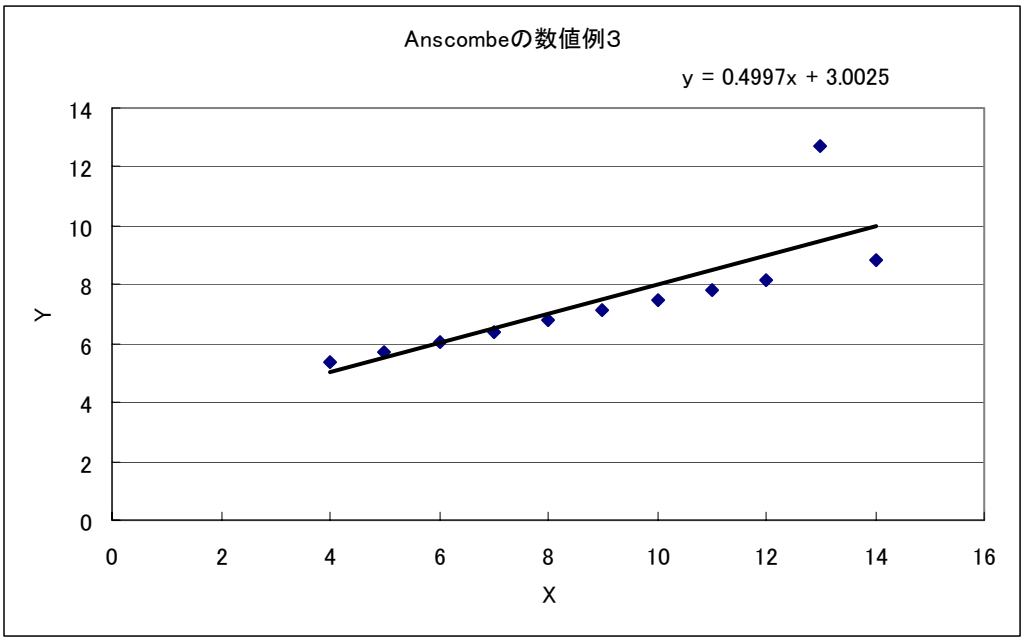


図 14　ほとんど直線で外れ値１つ

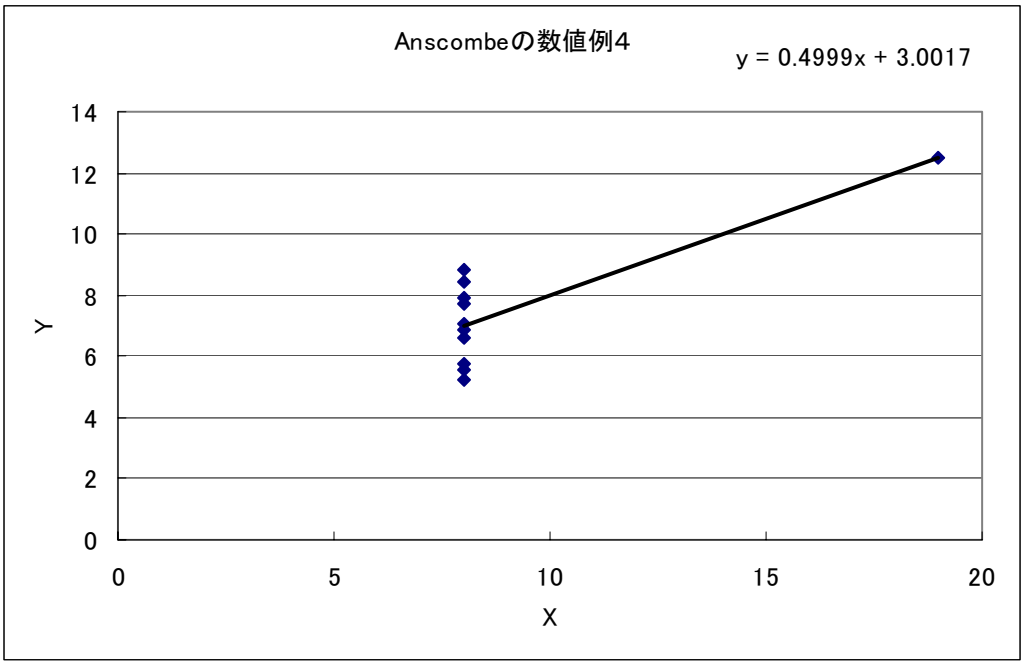


図 15　異常値によるみせかけの変動

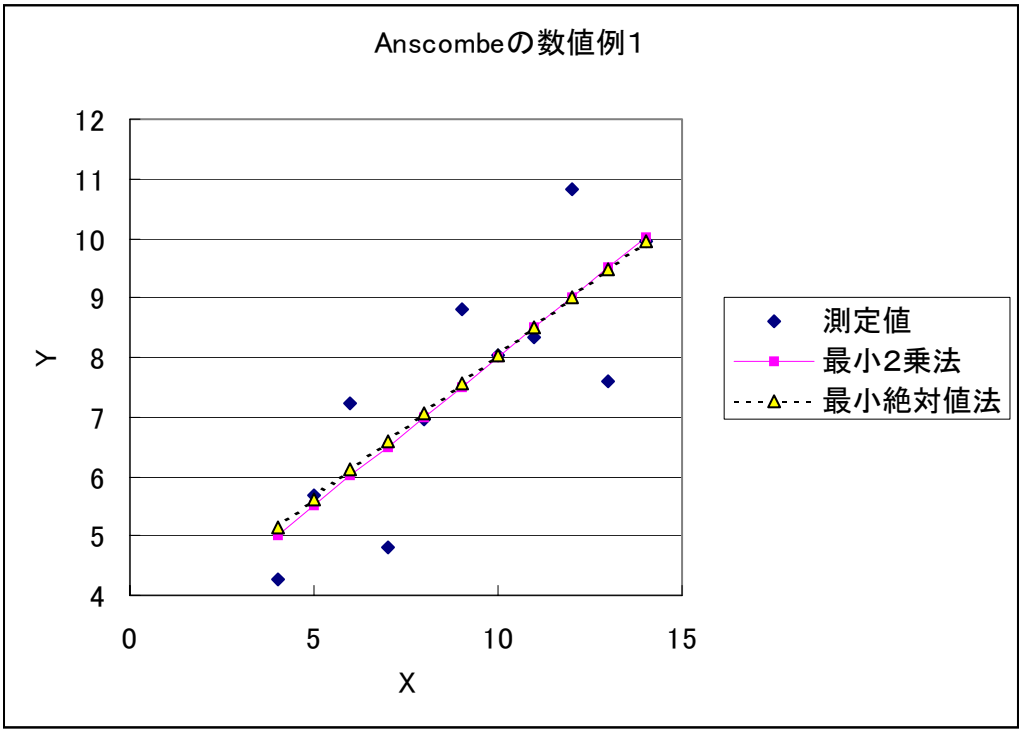


図 16 性質のいいデータでの推計方法比較

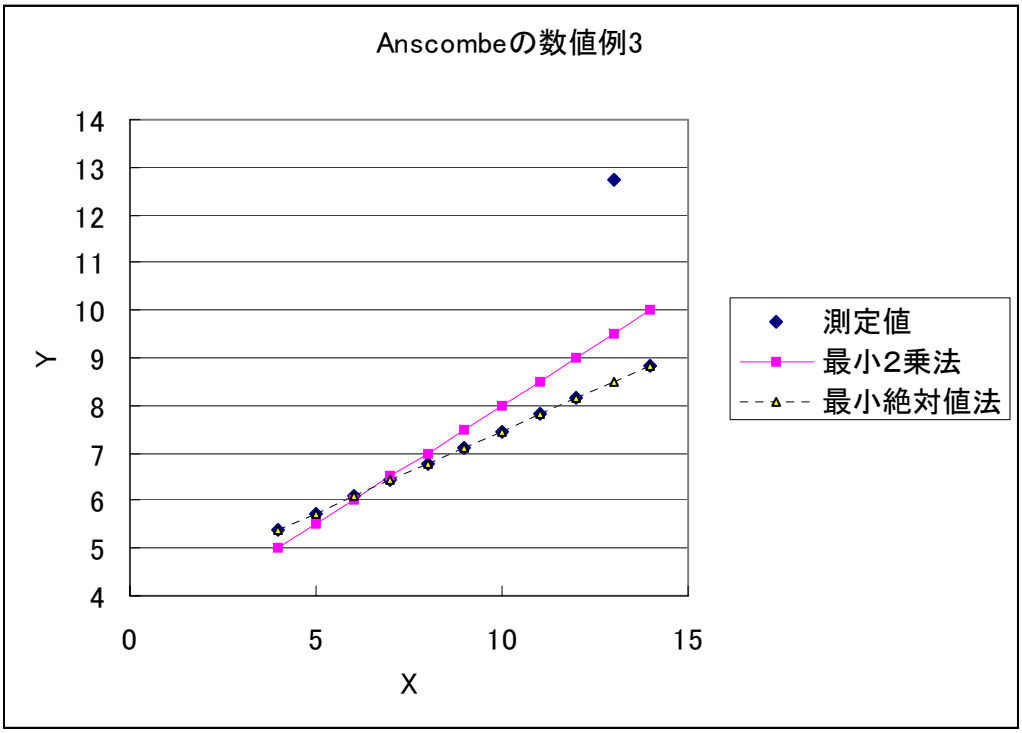


図 17　性質の悪いデータでの推計方法比較