

第 13 回 単回帰分析

A. 回帰

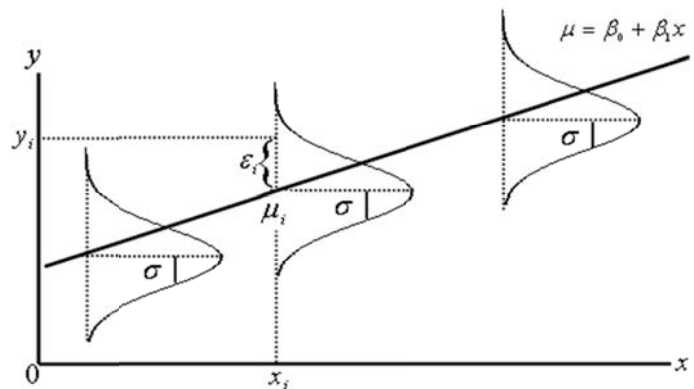
1. 回帰

x が指定した値であり, y が指定された x に対して, 誤差が正規分布にしたがいながら決定する場合, 両変数間の関係を回帰分析する. x と y の関係は直線だけでなく, 指数, 対数, 2 次曲線なども考えられる. ここでは直線の場合だけを考える (単回帰分析).

2. 回帰直線

x と y の 2 つの変数の関係を表した直線的な式を回帰直線という.

回帰直線のモデル (母回帰式) は右の図のようになっている. このことから単回帰分析の適用できる条件は基本的には以下の 2 つである.



- ① x が指定されたときに y は正規分布にしたがって決定する.
- ② x の値にかかわらず, y の標準偏差は一定である.

ただしこの条件が満たさないときでも変数変換などで単回帰分析できることも多い.

以上のようにして, 標本から求めた標本回帰式に対して, 母回帰式を考えることができる. 標本回帰式の傾き (回帰係数) と切片について推定・検定や分散分析をすることができる.

3. 回帰分析の基本用語

- ① () 変数 指定できる変数 x のこと. 独立変数ともいう.
- ② () 変数 y のこと. この値をコントロールすることが目的なので目的変数と呼ぶ. x の値が決まると y の母平均が決まるので, 従属変数ともいう.
- ③ 母回帰直線 母集団での回帰直線. この傾きを母回帰係数, 切片を母切片という. 両者をあわせて回帰母数という.
- ④ 予測値 x をある値にしたときに y がどのような値になるかを考えたものをいう. ある信頼区間をつけて予測値を求めることができる.
- ⑤ 残差 目的変数のデータの値と予測値との差.
- ⑥ 寄与率 SR/S_{yy} で求められる. また相関係数を二乗した値でもある. y の変動のうち回帰による変動の割合を示す.

B. 単回帰分析

1. 相関係数

回帰の時でも相関係数を同じように計算できる．ただし先週述べた母相関係数に関する推定や検定はできない．相関係数を二乗した値である寄与率の方が回帰では重要な値である．

2. 回帰係数

回帰係数は説明変数がある一定量変化したときに，目的変数がどれだけ変化するかを示す値であり，回帰直線で最も重要な値である．

回帰係数が 0 であるときはその両者の値の間に相関がない．

回帰係数が正（負）の時は，両者の値の間には正（負）の相関がある．

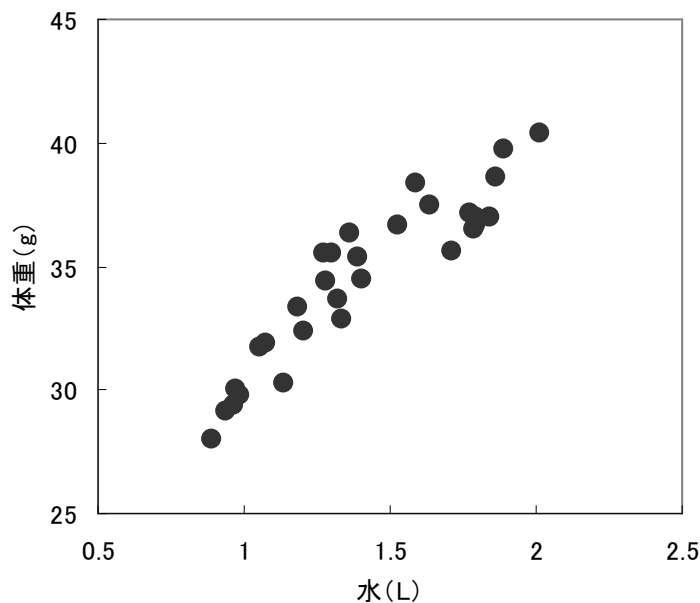
回帰直線 $\hat{y} = A + Bx$ はその傾きである回帰係数 B とその y 切片 A から成り立っている．

回帰直線は両変数の平均を座標とする点を通る．傾き B を回帰係数という（相関係数と混同しないこと）． y 切片の単位は目的変数と同じ，回帰係数 B の単位は目的変数／説明変数である．

3. 回帰分析の手順

例：与えた飲み水の量（単位：L，独立変数）によって，ウズラのヒナの体重（単位：g，従属変数）がどのように決定されるかを回帰分析した．

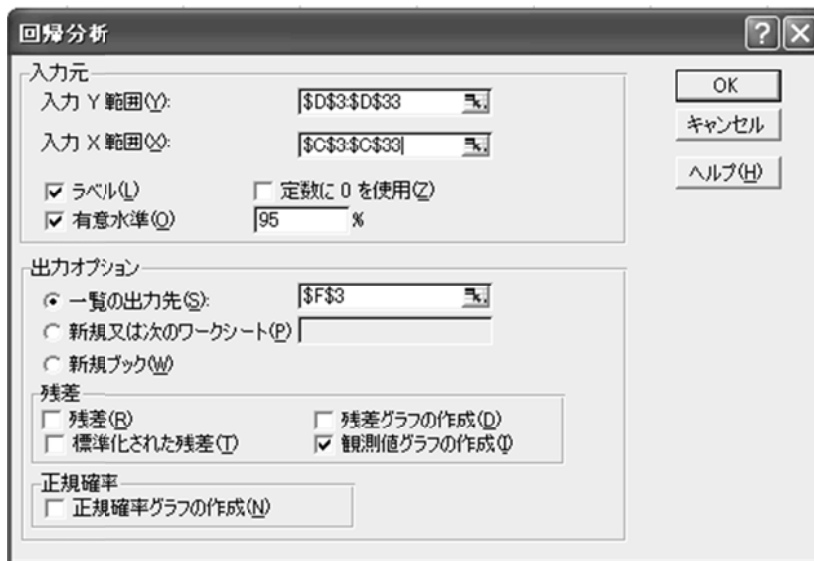
- ① 2 つの変数のうちどちらが独立変数であり，どちらが従属変数かを確認し，グラフを書く．



ウズラ	水(L)	体重(g)
1	1.864	38.625
2	1.775	37.125
3	1.794	37
4	1.84	37
5	1.792	36.625
6	1.71	35.625
7	1.392	35.375
8	1.334	32.875
9	1.403	34.5
10	1.32	33.625
11	1.785	36.5
12	1.635	37.5
13	1.89	39.75
14	2.013	40.375
15	1.077	31.875
16	0.984	29.75
17	0.97	30
18	0.964	29.375
19	1.526	36.625
20	1.589	38.375
21	1.363	36.375
22	1.281	34.375
23	0.935	29.125
24	0.887	28
25	1.054	31.75
26	1.185	33.375
27	1.273	35.5
28	1.299	35.5
29	1.14	30.25
30	1.205	32.375

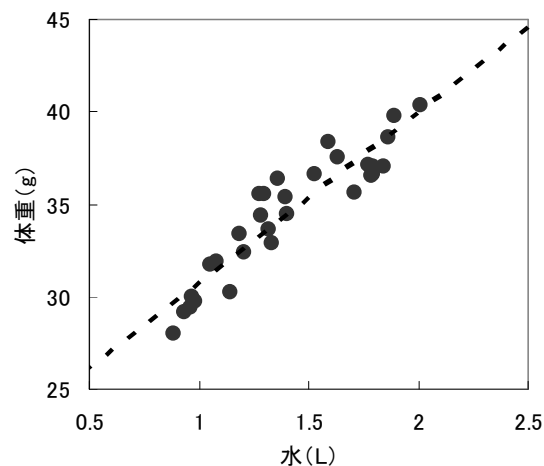
② 回帰係数，回帰直線の式を求める→③ 分散分析を行う

エクセルの分析ツールを使えば回帰直線の計算，分散分析などが一度にできる。



概要									
回帰統計									
重相関 R	0.926356								
重決定 R2	0.858136								
補正 R2	0.853069								
標準誤差	1.283244								
観測数	30								
分散分析表									
	自由度	変動	分散	測された分散	有意 F				
回帰	1	278.9071	278.9071	169.3718	2.15E-13				
残差	28	46.10803	1.646715						
合計	29	325.0151							
	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 99.0%	上限 99.0%	
切片	21.50633	1.025847	20.96446	1.18E-18	19.40498	23.60769	18.67165	24.34102	
水(L)	9.2229	0.708675	13.01429	2.15E-13	7.771246	10.67455	7.264646	11.18115	

エクセルの表から $\hat{y} = 21.51 + 9.22x$ という回帰直線の式が得られた。



単回帰分析の分散分析では要因を回帰と残差（偶然誤差だけでなく、回帰で採用した因子以外のすべての因子の影響も含んでいる）の 2 つに分け、回帰による変動が残差変動より十分に大きいかを F 検定する。

単回帰分析の分散分析表は以下のようにする。

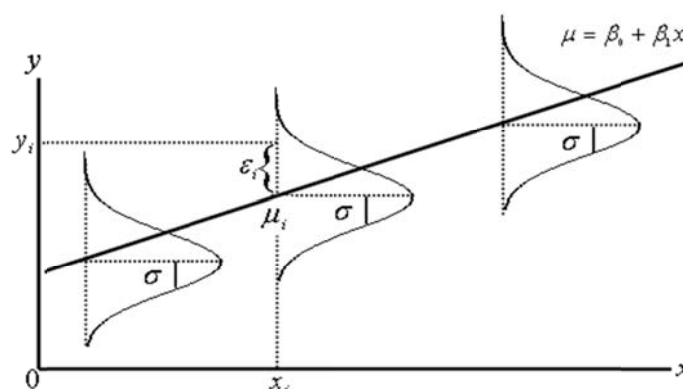
要因	平方和（変動）	自由度	平均平方（分散）	分散比	p-値
回帰	S_R	$\phi_R = 1$	$V_R = S_R / \phi_R$	V_R / V_e	p_R
残差	S_e	$\phi_e = n - 2$	$V_e = S_e / \phi_e$		
総変動	S_T	$\phi_T = n - 1$			

エクセルで出力された分散分析表から回帰の項を見て、有意 F（p-値：回帰がないという帰無仮説が成り立つとした場合、今回のようなデータの得られる確率である）が 0.05 や 0.01 より小さいと、5 %あるいは 1 %の有意水準で回帰は有意であると結論できる。

④ 分散分析の結果、回帰が有意であれば、必要に応じて、回帰係数などの推定や検定を行う

C. 回帰分析における推定と検定

回帰分析では独立変数 x を与えたときに従属変数 y がどんな値となるかを決めることが主な目的である。回帰係数、回帰直線などはそのために算出するのである。回帰係数、切片、独立変数 x に対する y の値について推定・検定ができる。



1. 回帰係数の推定と検定に必要な理論式

$\hat{y} = A + Bx$ という回帰式が標本から得られたとき、母集団の回帰式（母回帰式）を

$\mu = \alpha + \beta x$ と表せる。 α は母切片、 β は母回帰係数、誤差分散 V_e = 母分散 σ^2 である。

母回帰係数の分布は $N(\beta, \frac{\sigma^2}{S_{xx}})$ という正規分布に従う。

母切片の分布は $N(\alpha, (\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})\sigma^2)$ という正規分布に従う。

$\hat{\mu} = \hat{\alpha} + \hat{\beta}x_0$ の分布： $\hat{\mu}$ は $N(\alpha + \beta x_0, \{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\}\sigma^2)$ という正規分布に従う。

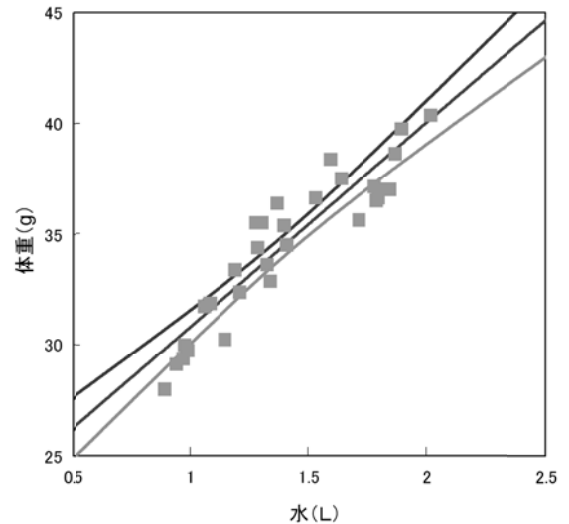
2. 回帰直線に対して、推定値はどのようにばらつくか？

回帰係数の推定と検定の計算は面倒であり、しかもパソコンのできるので、ここでは図を見て、データのばらつきがどのようなものを理解することに重点を置く。

右の図は先ほどデータから得られた回帰直線と y の平均値の 95%信頼区間を示している。

この図から次のことがわかる。

- ① 平均に近いほど平均の信頼区間は狭くなる．すなわち推定精度が高い．
- ② 平均から遠いほど、特に回帰式を求めるデータの範囲外にでると、信頼区間は広くなり、精度は落ちる．
- ③ 以上のことから、回帰分析では独立変数 x はできるだけ広い範囲をカバーすることが望ましい．さらに独立変数 x がカバーしない部分で y を推定するとあまり精度は高くないことがわかる．



3. 推定の誤差はどうすれば小さくなるか？

- ① 独立変数 x の範囲を広く取る
- ② 標本数を増やす
- ③ 平均に近いところを推定すると精度が高いから、推定したい x が平均付近に来るようにデータを集める

以上のことを母回帰係数と母切片の正規分布から説明すると以下のようなになる。

母回帰係数の分布は $N(\beta, \frac{\sigma^2}{S_{xx}})$ という正規分布に従う。

したがって、独立変数 x の範囲が大きくなるほど独立変数の平方和 S_{xx} は大きくなるから、母回帰係数の標準偏差は（大きく・小さく）なる。

母切片の分布は $N(\alpha, (\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})\sigma^2)$ という正規分布に従う。

したがって、母切片の標準偏差はそれぞれ独立変数 x の範囲が大きくなるほど（大きく・小さく）、サンプル数が増えるほど（大きく・小さく）、平均 $\bar{x}$ が 0 から遠くなるほど（大きく・小さく）なる。

$\hat{\mu} = \hat{\alpha} + \hat{\beta}x_0$ の分布： $\hat{\mu}$ は $N(\alpha + \beta x_0, \{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\}\sigma^2)$ という正規分布に従う。

したがって、 y の推定値の標準偏差はそれぞれ、独立変数 x の範囲が大きくなるほど（大きく・小さく）、サンプル数が増えるほど（大きく・小さく）、指定した x の値が平均 $\bar{x}$ に近いほど（大きく・小さく）なる。

4. 母回帰係数と母切片の推定・検定

回帰分析で独立変数 x が従属変数 y に対して何らかの効果を与えていることを検定するには母回帰係数が 0 であるという帰無仮説を立て、これを棄却できるかを検定すればよい。母回帰係数が 0 であるとは独立変数 x とは無関係に、従属変数 y が決まることを示す。

帰無仮説が棄却された場合

回帰直線の傾きが 0 でない。

独立変数 x を変化させると、 y が変化することを示す。

独立変数 x で表される因子は y に影響を与えることを示す。

相関の場合と同じく、回帰が有意であるからといって現実独立変数 x が y に直接的に効果を与えているとは限らない。

また、 p -値が小さいことと両変数間の関係が強いこととは関係がない。寄与率 $= \frac{S_R}{S_T}$ (エクセルでは重決定 R^2) によって独立変数がどれだけ従属変数を決定しているかを示すことができる。

帰無仮説が棄却できなかった場合

回帰直線の傾きが 0 でないとはいえない

独立変数 x で表される因子は y に影響を与えたとはいえない。

エクセルでは回帰および回帰係数の検定は有意 F (p -値と同じである) をみて、検定する。

概要									
回帰統計									
重相関 R	0.926356								
重決定 R2	0.858136								
補正 R2	0.853069								
標準誤差	1.283244								
観測数	30								
分散分析表									
	自由度	変動	分散	割された分散	有意 F				
回帰	1	278.9071	278.9071	169.3718	2.15E-13				
残差	28	46.10803	1.646715						
合計	29	325.0151							
係数		標準誤差	t	P-値	下限 95%	上限 95%	下限 99.0%	上限 99.0%	
切片	21.50633	1.025847	20.96446	1.18E-18	19.40498	23.60769	18.67165	24.34102	
水(L)	9.2229	0.708675	13.01429	2.15E-13	7.771246	10.67455	7.264646	11.18115	

母回帰係数の区間推定はエクセルで計算できる。エクセルでは 95% 信頼区間をつけて母回帰係数を区間推定できる。さらに右の画面の有意水準に p % を入力すれば、 p % 信頼区間をつけた母回帰係数の区間推定もできる。

?

×

入力元

☒ ラベル(L)
 ☐ 定数に 0 を使用(O)

☒ 有意水準(O)
 %

出力オプション

☒ 一覧の出力先(S):

☐ 新規又は次のワークシート(P)

☐ 新規ブック(W)

残差

☐ 残差(R)
 ☐ 残差グラフの作成(O)

☐ 標準化された残差(I)
 ☒ 観測値グラフの作成(O)

正規確率

☐ 正規確率グラフの作成(O)

母切片の推定と検定もできる．エクセルでは母切片が0であるという帰無仮説におけるp-値と95%信頼区間をつけた母切片区間推定（さらに p%信頼区間をつけた母切片の区間推定）が計算できる．なお、切片は $x=0$ のときの y の値であるから、 $x=0$ があり得ないデータについて、母切片の推定や検定をしても意味はない．例えば父と子の身長についての回帰式では、父の身長が0であることはあり得ない．この場合、切片について推定・検定はしても意味はない．

母回帰係数の95%信頼区間を付けた区間推定値 $7.77 \leq \beta \leq 10.67$ （単位 g/L）

母切片の95%信頼区間を付けた区間推定値 $19.40 \leq \alpha \leq 23.61$ （単位 g）

5. 回帰による推定と推定値の信頼区間

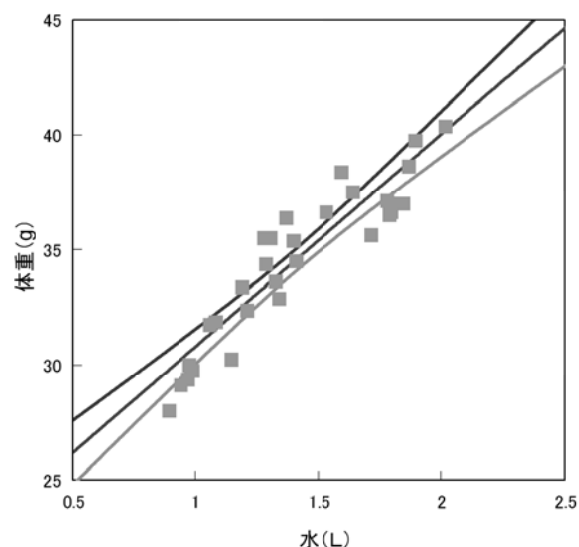
① η_0 の点推定

x_0 に対する η_0 の点推定は $\hat{\mu} = \hat{\alpha} + \hat{\beta}x_0$ の式から計算する．例えば、飲み水の量とウズラのヒナの体重の回帰の例で、飲み水 $x_0 = 1.5$ に対するウズラの体重 y の母平均 η_0 はエクセルでは以下のように回帰分析結果の回帰係数と切片を代入して計算すればよい．

平均	1.4093	34.50417	
平方和	3.278872	325.0151	
推定値	1.5	35.34068	=C39*G20+G19

② η_0 の区間推定

しかし、上の点推定のもととなる回帰式は標本から求めた標本回帰式であり、母回帰式は標本回帰式そのものではない．したがって、 y の平均値 η_0 を推定するときには誤差がふくまれるはずである． x_0 に対する η_0 の信頼区間の計算は難しいのでここでは説明しない．しかし、先述のウズラのヒナの例について計算した場合、グラフに x_0 に対する η_0 の95%信頼区間（点線と点線の間が信頼区間、実線は点推定）を図示すると右下の図のようになる．平均に近いほど推定の精度が高く、標本の範囲を超えると推定の精度が急速に落ちることがわかる．



D. 回帰分析における予測

1. 推定と予測の違い

回帰分析である x に対して y がどんな値になるかを知りたいときに、 y の平均値 η_0 を区間推定したいのではなく、ある 1 回だけ得られる $\hat{y}_0$ を区間推定したいときがある。例えば、薬 A を飲んで、ある患者の血圧がどの程度下がるかを知りたいときに、母平均（すなわち薬を飲んだすべての人の平均）の区間推定ではなく、そのたった一人の患者では 95% の信頼区間で血圧がどうなるかを知りたいだろう。そういう場合、このような予想をすることを予測という。推定は母数（母平均など）について使い、個々のデータについて統計的に予想するときには予測という言葉を用いる。

推定と予測の違い

天気予報は（推定・予測）である。したがって、明日の降水確率は 90% というときは、その明日 1 日だけについて（推定・予測）している。一方、過去 30 年の気象から 11 月 3 日は晴れの特異日であるというときは、降水確率を（推定・予測）していると考えることができる。

製薬会社 A としては風邪薬 Q が従来の薬 P より平均して効果があるかを知ろうとするから、新薬 Q の効果を（推定・予測）する。医師として新薬 Q が対象となる薬 P よりある患者 B に効果があるかを知ろうとするから新薬 Q の効果を（推定・予測）する。

研究者 C は温度が上昇すると、みかんの糖度がどうなるかを（推定・予測）した。みかん農家 D は温度が上昇すると、今年のみかんの糖度がどうなるかを（推定・予測）した。

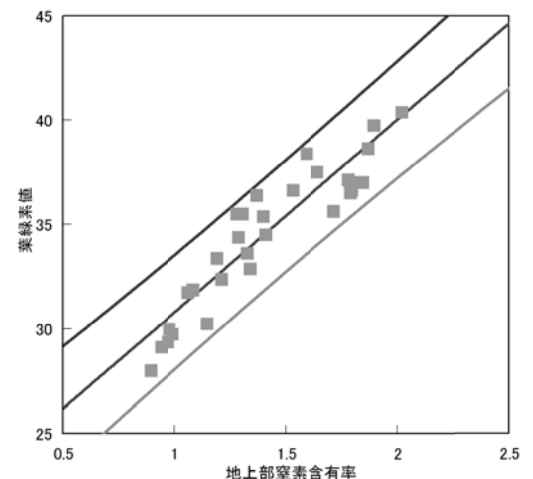
2. 予測の信頼区間

① y_0 の点予測

x_0 に対する y_0 の点予測は点推定 η_0 の場合と同じである。すなわち $y_0 = \eta_0$ である。

② 予測の信頼区間

予測のときの信頼区間の計算も面倒なので省略する。ここでは信頼区間をグラフに示して、推定と予測における信頼区間の違いをみる。先述のウズラのヒナの例について計算した場合、グラフに x_0 に対する y_0 の予測の 95% 信頼区間（点線と点線の間が信頼区間、実線は点予測）を図示すると右図のようになる。予測の場合は母平均の推定に比べると、データの平均に近づいても信頼区間の幅はあまり小さくならない。その理由は、回帰分析では独立変数 x に対して、従属変数 y がある幅のあるばらつきをとまって決まるという仮定のため、そのばらつきによって決まる誤差よりは信頼区間は小さくならないこと、さらにそのばらつきは x の値にかかわらず一定であると仮定しているからである。

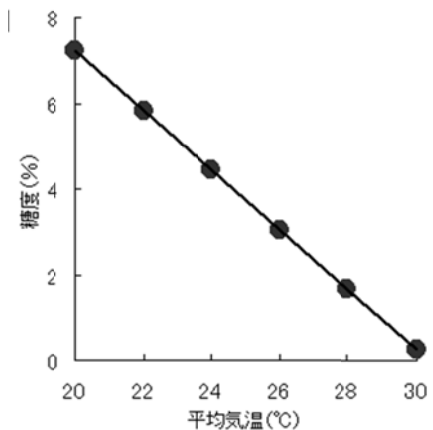


E. 回帰の逆推定

1. 逆推定とは？

回帰分析では独立変数 x は指定できる値であり，従属変数 y はあるばらつきをともなう値であるという前提で行う．次の例は先ほど用いた平均気温と果実の糖度の関係である．この関係では平均気温を指定すると果実の糖度が決まるという関係になっている．しかし，果実を生産する現場からすると，果実の糖度を何パーセント以上にするには平均気温を何度以下に設定したらよいかという問題の方が現実的である．では果実の糖度を独立変数 x にして，平均気温を y にして回帰分析したらよいのか？

平均気温	糖度
20	7.23
22	5.83
24	4.44
26	3.04
28	1.65
30	0.25



実は回帰分析では右の図のように x と y を入れ替えると得られる回帰式は平均を共通して通る異なる2つの直線になる．右の例は回帰分析には向かないデータであるが，仮に回帰分析できるとしたら1月の気温と7月の気温のどちらを独立変数とするかによって，得られる回帰式が異なることを右のグラフは示している．

回帰分析の前提条件から独立変数 x は指定できる値でなければならない．すなわち果実の例では独立変数 x は平均気温でなければならない．もし果実の糖度を指定した上で，平均気温を決めたいのであれば，平均気温を独立変数とした回帰式を求めてから， x についてこの式を解けばよい．

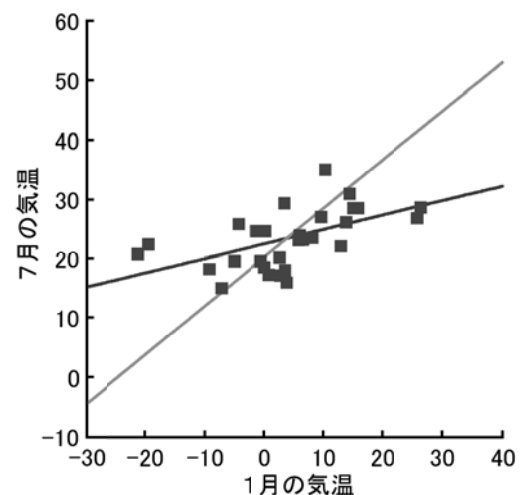
すなわち 果実の糖度 $y = 21.18286 - 0.69771 \times \text{平均気温}$ だから

平均気温 $x = -(\text{果実の糖度} - 21.18286) / 0.69771$ である．

このように従属変数 y から独立変数 x を回帰によって求めることを逆推定という．

化学分析で標準液をいくつか使って，検量線を書くのも逆推定である．標準液の濃度を指定すると吸光度などの反応量が決まる．未知試料の濃度は反応量から逆推定される．

回帰分析では独立変数 x は指定できる値を使うのが原則である．2つの分析法を比較する場合は両変数ともにばらつきがあるが，独立変数 x をよりばらつきの小さい分析法にすれば実用上問題ない．



例：先述のウズラのヒナの例で，ヒナの体重として 35g を得た．飲み水の量を逆推定せよ．

$$y = 21.50633 + 9.2229x$$

$$y = 35 \text{ を代入し, } x \text{ を求める. } x = (35 - 21.50633) / 9.2229 = 1.463061(\%)$$

逆推定	1.463061	= (F37 - G19) / G20
35		

2. 逆推定での信頼区間

逆推定での信頼区間の計算も煩雑である．ここでは専門的すぎるので，逆推定でも右の図のようにして信頼区間を計算できることだけを紹介するにとどめる．現実問題としては，果実の糖度の例でも果実の糖度が望む糖度より高くなる確率を 95% 以上にしたということがあるなら，信頼区間を計算しなければならないことになる．

