【セミナー】

線形回帰と相関

――社会調査データの多変量解析（1）――

小林　久高・山本　圭三
KOBAYASHI Hisataka, YAMAMOTO Keizo

1　はじめに

今回から多変量解析についてシリーズで解説する。まず、第 1 回目は回帰分析についてである。ここで取り扱うのは 1 つの独立変数から 1 つの従属変数を説明するような単回帰分析である。独立変数が 1 つなので、まだ多変量の解析とはいえない。しかしながら、これがわかっていないと重回帰分析などは理解できない。多変量解析の基礎はここに詰まっているのだ。

回帰分析には多くのテキストがあり、その中にはとてもよくできたものもある。しかしながら、解説が簡単すぎてぼやっとしかわからないものや、難しすぎて途中で投げ出したくなるものも多い。ちょうどいいものが少ないなあ、というのが筆者らの印象である。そこで、今回は 1 つの工夫をした。工夫とは、今回の解説を 5 つの部分から構成したということだ。

本セミナーでは、2 節でまず平均、分散、偏差平方和、偏差積和、相関、共分散など、基本的な統計量についておさらいをする。次の 3 節では、記述統計の世界の中で回帰分析を解説する。ここでは、母集団と標本などということを考えず、得られたデータの世界だけで回帰分析を考えてみるのである。次の 4 節では、ベクトルを導入して回帰分析を図形的にとらえてみる。決定係数や相関係数をイメージとして感覚的に把握できるようにするのがこの節の役割だ。次の 5 節では、母集団と標本という考えを導入し、回帰分析を推測統計的に解説する。この節では推測統計的な回帰分析の意味がストレートにわかるように、できるだけ枝葉の解説はしていない。そのほうが重要な点をきちんと理解するためには都合がいいと考えたからである。この解説によって、統計ソフトなどで算出される結果の表の意味がすべてわかるようになると思う。最後の 6 節では 5 節の議論がなぜ成り立つかを数学的に解説する。なぜ推測統計的な回帰分析のロジックが成り立つのだろうと思う読者のために書かれている。

このような構成をとっているので、本セミナーでは、途中のどこで読むのをやめてもそれなりの知識は得られる。回帰分析とは何かということをざっと知りたい読者は 3 節まで読めばいい。この分析のイメージをさらに豊かにしたければ 4 節まで読むといい。回帰分析の検定や推定まで知りたければ 5 節まで読めばいい。検定や推定の数学的根拠まで知りたい読者は最後の 6 節まで読む必要がある。

5 節と 6 節は推測統計に関わるため、確率変数と確率分布の知識が必要である。本文の中でも基礎的な事項はまとめておいたが、小林（2018a; 2018c）と小林（2019a）の 2 節と 3 節、ならびに小林（2019b）の補 1 を読んでおいたほうが理解が早いと思う。

2　準備

2.1 基本的な統計量

（1）平均・分散・標準偏差

1）平均

n 個の値があるとき、その分布の平均は次式で表される。平均はデータの分布を代表する値だ。

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\Sigma x_i$$

母集団の平均は μ で表される。

2）分散

n 個の値があるとき、その分布の分散は次式で表される。分散はデータの散らばりを表す値だ。

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

それが x の分散であることを示したいときには、添え字がつけられる。

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

分散は大文字の V で示されることもある。

$$V_x = s_x^2 = \frac{1}{n}\sum\left(x_i - \bar{x}\right)^2$$

分散は母集団では σ^2 で表される。

3）不偏分散

n 個の値があるとき、その分布の不偏分散は次式で表される。

$$u^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$$

この式での分散は推測統計において用いられる。不偏分散を s^2 や V_x で表すことも多い。その場合は次のようになる。

$$V_x = s_x^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$$

分散という用語は、分散を示すこともあれば不偏分散を示すこともある。s^2 という記号があるとき、分散を示しているのか、不偏分散を示しているのかについては十分注意が必要だ。

4）標準偏差

分散の正の平方根を標準偏差という。

$$s_x = \sqrt{s_x^2} = \sqrt{V_x}$$

標準偏差は n で割った分散をもとにしているか、n－1 で割った分散をもとにしているかで値が変わる。

変数が cm で測られた身長だとすると、分散は cm^2 の単位、標準偏差は cm の単位である。標準偏差はもとの変数の単位と同じだから、変数の値と関連づけて考えやすい。標準偏差は母集団では σ で表される。

（2）偏差平方和と偏差積和

1）偏差平方和（平方和・変動）

n 個の値があるとき、下式で算出される値は偏差平方和（平方和・変動）といわれる。

$$S_{xx} = \sum(x_i - \bar{x})^2$$

$$= \sum x_i^2 - n\bar{x}^2$$
$$= \sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2$$

大文字のＳであることに注意。小文字の s は標準偏差である。偏差平方和を n または n−1 で割ると分散が出る。偏差平方和はデータの散らばりの 1 つの指標である。

2）偏差積和

ｎ個の値があるとき、下式で算出される値は偏差積和といわれる。偏差平方和と同様、大文字のＳであることに注意してほしい。

$$S_{xy} = \sum_1^n \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$
$$= \sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i$$

(3) 共分散と相関係数

1）共分散

共分散（Cov）は偏差積和を n または n−1 で割ったものだ。

$$Cov(x,y) = s_{xy} = \frac{S_{xy}}{n}$$
$$Cov(x,y) = s_{xy} = \frac{S_{xy}}{n-1}$$

分散や共分散は小文字の s を使い、偏差平方和や偏差積和は大文字の S を使うことに注意が必要だ。偏差平方和や偏差積和が n（あるいは n−1）で割る前の世界に属し、分散や共分散が n（あるいは n−1）で割ってからの世界に属すこと、標準偏差がそのルートの世界に属することをきちんと押さえておく必要がある。母集団の共分散は σ_{xy} と表わされることもある。

2）相関係数

相関係数は次の式で定義される。

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{Cov(x,y)}{\sqrt{V_x V_y}}$$
$$= \frac{s_{xy}}{s_x s_y} = \frac{\frac{1}{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\frac{1}{n}\left(x_i - \bar{x}\right)^2}\sqrt{\frac{1}{n}\left(y_i - \bar{y}\right)^2}}$$

n で割った分散や共分散を使っても n−1 で割った分散や共分散を使ってもこの値は同じになる。相関係数は偏差積和と偏差平方和を用いて、上のように簡単に表現できる。

相関係数は量的 2 変数の直線的関係の大きさを表す指標である。範囲は−1〜＋1 をとり、一方が大きくなれば他方も大きくなる関係があるとき＋1 に近づく（正の相関）。一方が大きくなれば他方が小さくなる関係があるとき−1 に近づく（負の相関）。このような関係がないとき 0 になる。大きい相関とは＋1 や−1 に近い相関のことをいう。相関係数には平均や標準偏差と異なり単位はない（こういった数を無名数という）。母集団の相関係数は ρ（ロウ）で表されることがある。

2.2 変数と得点の変換

(1) 素点・平均偏差得点・標準得点

表 1 はある仮想的な調査で得られたデータである。得られたデータは素点の列に並べられている。社会調査の解析ではこのデータをいろいろな形に変換して分析を行う。

平均偏差得点とは素点から平均を引いた得点だ。たとえば 1 氏の場合、素点の 28 から素点の平均

である 39.6 を引いた－11.6 が 1 氏の平均偏差得点になる。素点を平均偏差得点に変えたとき、平均偏差得点の平均は 0 になるが、分散や標準偏差は変わらない。

表 1　素点・平均偏差得点・標準得点

	年齢		
	素点	平均偏差得点	標準得点
1氏	28	-11.6	-1.36
2氏	35	-4.6	-0.54
3氏	36	-3.6	-0.42
4氏	49	9.4	1.10
5氏	50	10.4	1.22
n	5	5	5
平均	39.6	0.0	0.00
分散	73.0	73.0	1.00
標準偏差	8.5	8.5	1.00
偏差平方和	365.2	365.2	5.00

データ全体の標準偏差が 1（すなわち分散が 1）になるようにさらに値を変換したものが標準得点だ。標準得点は平均偏差得点を標準偏差で割ることによって得られる。1 氏の標準得点は、－11.6÷8.5＝－1.36 となる。ここでは平均が 0、分散は 1、標準偏差は 1 となっている。

素点から考えると、標準得点は平均を引き標準偏差で割るという作業で得られる。もとの得点を x_i で表すと、標準得点 x_i' は次のようになる。

$$x_i' = \frac{x_i - \bar{x}}{s_x} = \frac{x_i - \bar{x}}{\sqrt{V_x}}$$

（2）変数の標準化

変数の値をすべて標準得点に変換するということは、変数自体を標準化することでもある。もとの変数を x で表すと、標準化された変数 x' は次のようになる。

$$x' = \frac{x - \bar{x}}{s_x} = \frac{x - \bar{x}}{\sqrt{V_x}}$$

正規分布は平均 μ、分散 σ^2 をつかって $N\left(\mu, \sigma^2\right)$ と表わされるのだが、この正規分布に従う変数を標準正規分布 $N\left(0,1\right)$ に従う変数に変換することはよくある。ここで行われていることは、下に示すように、変数の標準化、すなわち、平均 μ を引き標準偏差 σ で割るということである（～は､左の変数は右の分布に従うということ）。

$$x \sim N\left(\mu, \sigma^2\right) \quad \Rightarrow \quad \frac{x-\mu}{\sigma} \sim N\left(0,1\right)$$

2.3　課題

（1）問題

a)分散と偏差平方和の関係、ならびに共分散と偏差積和の関係を式で示せ。
b)相関係数を共分散と分散で表せ。また偏差積和と偏差平方和で表せ。

（2）解答

a)

$$\frac{1}{n} S_{xx} = s_x^2 \qquad \frac{1}{n} S_{xy} = s_{xy}$$

n は n－1 の場合もある。

b)

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2}\sqrt{s_y^2}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

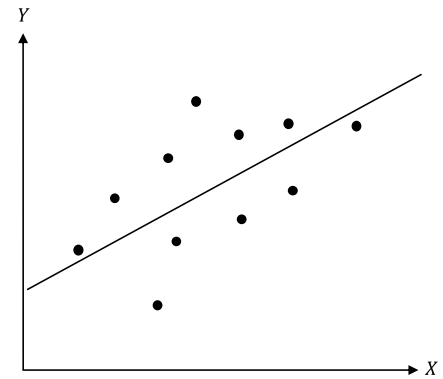
3　回帰分析の記述統計的理解

3.1 回帰直線

（1）データと予測直線

ある変数 x の値 x_i によって、別の変数 y の値 y_i の値が変化するようなことがある。たとえば、勉強時間 x が長いほど試験の成績 y がいいというのはこういったことだ。このとき、x を説明変数（独立変数）、y を被説明変数（従属変数）という。回帰分析とは、説明変数と被説明変数の間に直線的な関係を見立てて分析する方法である（図 1）。

図 1　データの散らばりと予測直線



（2）予測式と回帰直線の式

各 y_i の予測値の点 $\hat{y}_i$ を示す式は次のように表せる。ここではこれを回帰の予測値の式、あるいは予測式ということにする（「^」は予測値の意味）。

$$\hat{y}_i = b_0 + b_1 x_i$$

これを XY 軸のグラフに表記できるように、次のように書き直したものをここでは回帰直線の式と呼ぶ。

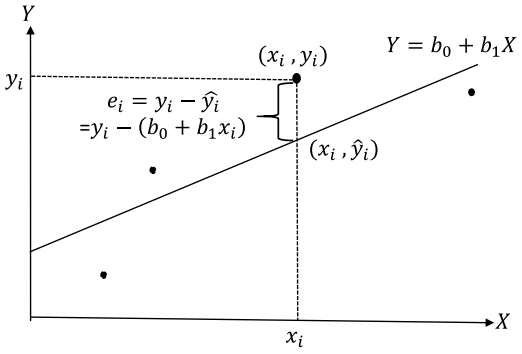
$$Y = b_0 + b_1 X$$

回帰直線の式は、x_i と y_i の関係を直線で表したものである。b_1 は回帰係数と呼ばれ、この直線の傾きを表す。b_0 は定数項であり、切片と呼ばれる。

（3）回帰係数と切片の算出

b_0、b_1 は最小二乗法という方法で求める。データの x_i に対応する値が y_i であるとき、予測値を $\hat{y}_i = b_0 + b_1 x_i$ とすると図 2 のような状況になっている。

図 2　残差

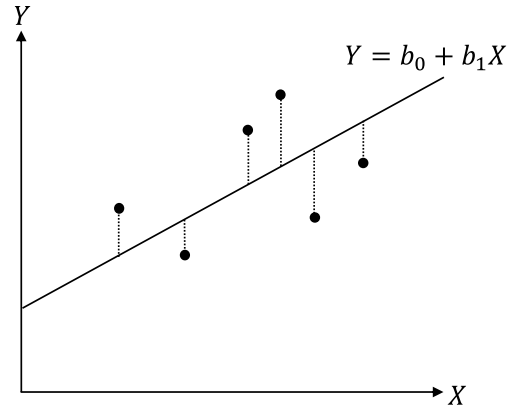


ここでデータの値と予測値の差「$y_i - \hat{y}_i$」、すなわち「$y_i - (b_0 + b_1 x_i)$」は残差 e_i と呼ばれるものだ。この残差の 2 乗（平方）をすべてのデータについて出し、それらの総和である残差平方和 $\sum_{i=1}^{n} \left(y_i - (b_0 + b_1 x_i) \right)^2$ を最小にするように、切片 b_0 と傾き b_1 を求めるのである（図 3）。

b_0 の値が大きくなりすぎても小さくなりすぎてもこの残差平方和の値は大きくなってしまう。b_1 の値が大きくなりすぎても小さくなりすぎても残差平方和の値は大きくなってしまう。b_0、b_1 を絶妙に組み合わせて「実現値と予測値の差の 2

乗和」を最小にするのである。

図 3　残差と回帰直線の傾き・切片



数学的にこの解を求めるためには、残差の平方和をb_0、b_1の関数と考えて、

$$f\left(b_0, b_1\right) = \sum \left(y_i - (b_0 + b_1 x_i)\right)^2$$

とおき、これをb_0、b_1で偏微分して、それぞれの導関数を 0 とすればいい。この計算を進めると、b_0とb_1は x、y の平均、偏差積和S_{xy}、偏差平方和S_{xx}で次のように表せる。

$$\begin{cases} b_1 = \dfrac{S_{xy}}{S_{xx}} \\ b_0 = \bar{y} - \dfrac{S_{xy}}{S_{xx}} \bar{x} \end{cases}$$

$\hat{y}_i = b_0 + b_1 x_i$なので、予測値の式はこうなる。

$$\boxed{\hat{y}_i = \left(\bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x} \right) + \frac{S_{xy}}{S_{xx}} x_i}$$

これを回帰直線の式で表現すると次のようになる。

$$\boxed{Y = \left(\bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x} \right) + \frac{S_{xy}}{S_{xx}} X}$$

予測値の式や回帰直線の式は共分散s_{xy}や分散s_x^2を使っても表せる。

$$\hat{y}_i = \left(\bar{y} - \frac{s_{xy}}{s_x^2} \bar{x} \right) + \frac{s_{xy}}{s_x^2} x_i$$

$$Y = \left(\bar{y} - \frac{s_{xy}}{s_x^2} \bar{x} \right) + \frac{s_{xy}}{s_x^2} X$$

ここで、

$\hat{y}_i = b_0 + b_1 x_i$

という予測値の式は x や y の平均を使って、

$$\boxed{\hat{y}_i = b_1 (x_i - \bar{x}) + \bar{y}}$$

と表現できることも押さえておこう。

というのは、

$$\hat{y}_i = \left(\bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x} \right) + \frac{S_{xy}}{S_{xx}} x_i$$

$$\hat{y}_i = \left(\bar{y} - b_1 \bar{x} \right) + b_1 x_i$$

$$\hat{y}_i = \left(b_1 x_i - b_1 \bar{x} \right) + \bar{y}$$

$$\hat{y}_i = b_1 \left(x_i - \bar{x} \right) + \bar{y}$$

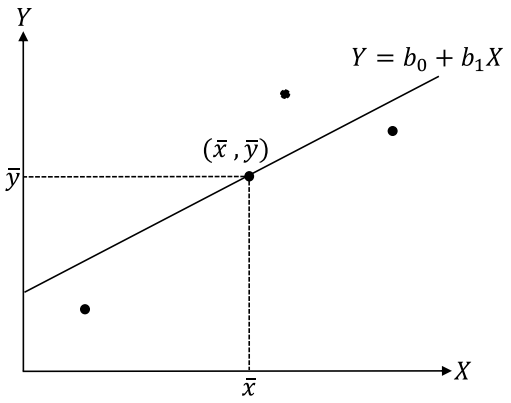
だからである。

x_i に$\bar{x}$を代入すると$y_i = \bar{y}$ となることから、次の式も成り立つ。

$$\boxed{\bar{y} = b_0 + b_1 \bar{x}}$$

$Y = b_0 + b_1 X$ という回帰直線は平均が交差する点 $(\bar{x}, \bar{y})$ を必ず通るということがここからわかる（図 4）。

図 4　回帰直線と平均の交差点



3.2 相関と回帰

（1）相関係数

x と y との相関係数は x と y との共分散を、それぞれの標準偏差で割ったものだった。それは x と y との偏差積和をそれぞれの偏差平方和のルートで割ったものともいえる。

$$r = \frac{s_{xy}}{s_x s_y} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

（2）相関が 0 の場合の回帰直線

相関係数 r が 0 のとき共分散 s_{xy} が 0 になるので、下のように回帰係数も 0 になる。このとき、

$$\hat{y}_i = \left(\bar{y} - \frac{s_{xy}}{s_x^2}\bar{x}\right) + \left(\frac{s_{xy}}{s_x^2}\right)x_i$$

$$= \left(\bar{y} - \frac{0}{s_x^2}\bar{x}\right) + \left(\frac{0}{s_x^2}\right)x = \bar{y}$$

となるから、

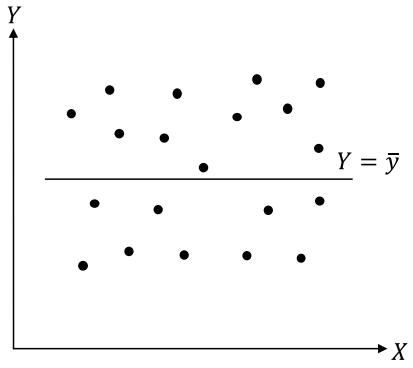
$$\hat{y}_i = \bar{y}$$

したがって、x から y を予測する回帰直線は、

$$Y = \bar{y}$$

となる（図 5）。

図 5　相関が 0 の場合の回帰直線



（3）相関係数と回帰係数

相関係数 r は次の式から求められた。

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{xy}}} = \frac{s_{xy}}{s_x s_y}$$

一方、回帰係数 b_1 は次の式から求められた。

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{s_{xy}}{s_x^2}$$

したがって、相関係数 r と回帰係数 b_1 の間には次の関係があることがわかる。

$$r=\left(\frac{s_x}{s_y}\right)b_1$$

ここから、x と y の標準偏差が同じ場合、すなわち分散が同じ場合、相関係数 r と回帰係数b_1は同じになることがわかる。

回帰係数b_1は回帰直線の傾きであり、x の増加に伴う y の増加を示すものだった。したがって等しい標準偏差をもつ x と y については、相関係数は、x が 1 単位増加すると y が何単位増加するかを示すものということになる。

（4）標準回帰係数

平均が 0 になるように変数を変換することがある。すなわち元の変数を平均偏差得点からなる変数に変換するのである。x と y 両変数がこのように変換されているとき、回帰の予測式は、

$$\hat{y}_i=\left(\bar{y}-\frac{S_{xy}}{S_{xx}}\bar{x}\right)+\frac{S_{xy}}{S_{xx}}x_i$$
$$=\left(0-\frac{S_{xy}}{S_{xx}}0\right)+\frac{S_{xy}}{S_{xx}}x_i=\frac{S_{xy}}{S_{xx}}x_i$$

となり、切片はなくなる（0 になる）。ただし、回帰係数の値はもとの回帰係数と同じだ。

変数をさらに、平均 0、分散 1 になるように変換することがある。これを変数の標準化という。x と y 両変数がこのように変換されているとき、回帰の予測式の切片はなくなるとともに、回帰係数の値も元の回帰係数の値と異なるものになる。標準化された変数の回帰係数のことを標準回帰係数という。

変数が標準化されているとき、次の式が成り立つ。

$$b_1=\frac{s_{xy}}{s_x^2}=\frac{s_{xy}}{s_xs_x}=\frac{s_{xy}}{s_xs_y}=r$$

すなわち、標準回帰係数は相関係数と等しくなるのである。

（5）y から x を予測する回帰直線

通常の回帰分析では、x から y を予測する。この回帰直線の式は次のようなものだった。

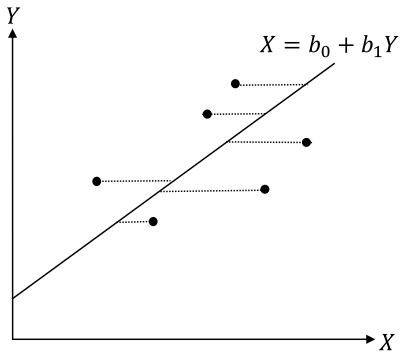
$$Y=\left(\bar{y}-\frac{S_{xy}}{S_{xx}}\bar{x}\right)+\left(\frac{S_{xy}}{S_{xx}}\right)X$$

ここで、同じロジックを適用し、y から x を予測する回帰直線の式を考えると次のようになる。

$$X=\left(\bar{x}-\frac{S_{xy}}{S_{yy}}\bar{y}\right)+\left(\frac{S_{xy}}{S_{yy}}\right)Y$$

これは横方向（x 軸方向）のデータの予測値と実測値のズレを最小にするような最小二乗法で求められた回帰直線である（図 6）。

図 6　y から x を説明する回帰直線



これを Y について解くと次のようになる。

$$Y=\left(\bar{y}-\frac{S_{yy}}{S_{xy}}\bar{x}\right)+\left(\frac{S_{yy}}{S_{xy}}\right)X$$

これは 2 つ上の式とは異なるので、x から y を予測する回帰直線と y から x を予測する回帰直線は、通常異なるということがわかる。

相関係数 r が 1 または−1 のとき 2 つの回帰直線は等しくなる。相関が 0 のとき、この回帰直線は、$X=\bar{x}$ となる。

3.3 平方和と決定係数

（1）全平方和・回帰平方和・残差平方和

回帰分析における重要概念として、全平方和、残差平方和、回帰平方和というものがある。全平方和（S_{yy}）とは、従属変数の各ケースの値 y_i と平均 $\bar{y}$ との差の平方和のことだ。これはデータの散らばりを表す指数である。

$$S_{yy}=\sum\left(y_i-\bar{y}\right)^2$$

残差平方和（S_e）とは、従属変数の各ケースの値 y_i と回帰の予測式による予測値 $\hat{y}_i$ との差（残差）e_i の平方和のことで、回帰の予測式では説明されない部分の平方和を意味する。

$$S_e=\sum\left(y_i-\hat{y}_i\right)^2$$

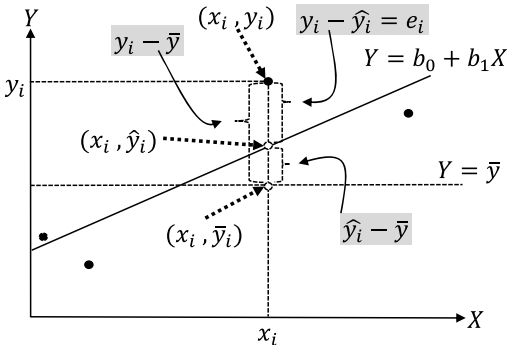
回帰平方和（S_R）とは、従属変数の各ケースの値についての予測値 $\hat{y}_i$ と平均 $\bar{y}$ との差の平方和のことだ。

$$S_R=\sum\left(\hat{y}_i-\bar{y}\right)^2$$

図 7 には $\left(y_i-\bar{y}\right)$、$\left(y_i-\hat{y}_i\right)$、$\left(\hat{y}_i-\bar{y}\right)$ が示されている。データ全体についてのこれらそれぞれの 2 乗の総和が、全平方和（全変動）、残差平方和（残差変動）、回帰平方和（回帰変動）である。

$$S_{yy}=\sum\left(y_i-\bar{y}\right)^2$$
$$S_e=\sum\left(y_i-\hat{y}_i\right)^2$$
$$S_R=\sum\left(\hat{y}_i-\bar{y}\right)^2$$

図 7　回帰直線と平均値・予測値・残差



（2）平方和に関する等式

最小二乗法で回帰係数や切片が求められている場合、平方和についての次の式が成り立つ。

$$S_{yy}=S_R+S_e$$
$$\sum\left(y_i-\bar{y}\right)^2=\sum\left(\hat{y}_i-\bar{y}\right)^2+\sum\left(y_i-\hat{y}_i\right)^2$$

左辺はデータの全平方和を意味する。右辺 2 項目は回帰の予測式で説明されなかった残差の平方和だ。したがって、右辺第 1 項は説明された平方和ということになる。そこで、これを T（total）を

使って、次のようにも表現することもある。

$$S_T = S_R + S_e$$
$$S_T \left(= S_{Total}\right) = S_{yy}$$

（3）決定係数

回帰の予測式がどの程度データの散らばりを説明するかに際して、決定係数 R^2 が求められることがある。式は次の通りだ。

$$R^2 = 1 - \frac{S_e}{S_{yy}} = \frac{S_R}{S_{yy}}$$

式からすぐわかるように、決定係数はデータの変動を回帰の予測式がどの程度説明するかを示すものだ。範囲は 0〜1 をとる。1 のときは予測式で説明しつくされていることを示し、0 のとき、その式では何も説明できないことを示している。

上では平方和 S で考えたが、分散 s^2 で考えても理屈はまったく同じである。

$$R^2 = \frac{S_R}{S_{yy}} = \frac{\frac{1}{n}S_R}{\frac{1}{n}S_{yy}} = \frac{s_R^2}{s_{yy}^2}$$

すなわち、決定係数はデータの分散を回帰の予測式がどの程度説明するかを示す。独立変数が 1 つの単回帰分析の場合、決定係数は相関係数の 2 乗になる。

3.4 回帰分析の実例 1

（1）データ

さて、以上の知識を前提として、ここで実際に回帰分析を行ってみよう。

表 2　自殺・高齢化・失業・所得（県別データ）

都道府県	自殺率	高齢化率	失業率	県民所得
1 北海道	19.5	29.1	3.5	2.589
2 青森県	20.5	30.1	4.2	2.462
3 岩手県	23.3	30.4	2.9	2.760
4 宮城県	17.4	25.7	3.7	2.987
5 秋田県	25.7	33.8	3.5	2.420
6 山形県	21.7	30.8	2.7	2.677
7 福島県	21.6	28.7	3.1	2.941
8 茨城県	18.6	26.8	3.2	3.079
9 栃木県	19.5	25.9	3.0	3.481
10 群馬県	21.6	27.6	2.8	3.145
11 埼玉県	18.0	24.8	3.2	2.977
12 千葉県	19.3	25.9	3.0	2.920
13 東京都	17.4	22.7	3.6	5.378
14 神奈川県	16.8	23.9	3.3	2.986
15 新潟県	22.0	29.9	2.9	2.778
16 富山県	20.5	30.5	2.5	3.373
17 石川県	18.3	27.9	2.3	2.949
18 福井県	15.4	28.6	1.8	3.196
19 山梨県	16.8	28.4	2.8	2.785
20 長野県	18.2	30.1	2.7	2.927
21 岐阜県	18.8	28.1	2.3	2.755
22 静岡県	18.7	27.8	2.7	3.316
23 愛知県	16.0	23.8	2.5	3.677
24 三重県	19.0	27.9	2.2	3.556
25 滋賀県	17.4	24.2	2.2	3.058
26 京都府	16.5	27.5	3.3	2.942
27 大阪府	18.7	26.1	4.2	3.127
28 兵庫県	17.6	27.1	3.8	2.752
29 奈良県	15.9	28.7	3.2	2.494
30 和歌山県	19.2	30.9	2.4	2.738
31 鳥取県	18.2	29.7	2.4	2.249
32 島根県	22.9	32.5	2.6	2.647
33 岡山県	18.2	28.7	3.0	2.744
34 広島県	17.5	27.5	3.0	3.074
35 山口県	20.0	32.1	2.8	2.774
36 徳島県	17.2	31.0	3.0	2.921
37 香川県	16.2	29.9	2.8	2.925
38 愛媛県	19.3	30.6	2.8	2.535
39 高知県	15.7	32.8	3.0	2.532
40 福岡県	17.8	25.9	4.1	2.724
41 佐賀県	16.6	27.7	3.0	2.412
42 長崎県	16.9	29.6	3.2	2.388
43 熊本県	19.9	28.8	3.5	2.438
44 大分県	16.5	30.4	2.9	2.619
45 宮崎県	23.2	29.5	3.2	2.315
46 鹿児島県	19.0	29.4	3.5	2.384
47 沖縄県	20.7	19.6	5.1	2.166

※データはすべて2015年度の値。

用いるデータは2015年の各都道府県の自殺率、高齢化率、完全失業率、1 人当たりの県民所得についてのものである（表 2)。指標の定義とデータの出典はそれぞれ以下のとおり。

・自殺率…人口 10 万人当たりの自殺者数（出典：平成 27 年 人口動態調査)。
・高齢化率…都道府県人口における 65 歳以上の者の割合（出典：平成 29 年版高齢社会白書)。
・完全失業率…労働力人口のうち完全失業者が占める割合（出典：平成 27 年 労働力調査)。
・1 人当たり県民所得…県民所得（企業所得、財産所得、雇用者報酬の合計）を各都道府県の人口で割ったもの（単位：百万円、出典：統計でみる都道府県のすがた 2019)。

（2）各変数の基本統計量と相互関係

具体的な回帰分析を行う前に、上にあげたデータについての基本統計量ならびに各変数の相互関係について示しておこう（表 3、表 4)。

表 3　各変数の基本統計量

	自殺率	高齢化率	失業率	県民所得
平均	18.845	28.285	3.051	2.874
偏差平方和	237.836	352.220	16.837	11.810
分散	5.060	7.494	0.358	0.251
標準偏差	2.250	2.738	0.599	0.501

表 4 から、正の相関が読み取れるのは、自殺率と高齢化率、自殺率と完全失業率の間である。また、負の相関が読み取れるのは、自殺率と県民所得､高齢化率と完全失業率､高齢化率と県民所得、完全失業率と県民所得の間である。高齢化率や失業率が高いほど自殺率は高くなり、県民所得が多くなるほど自殺率は低くなる。高齢化率が高くなるほど失業率や県民所得が低くなる、といった傾向があるようだ。高齢化率と県民所得の関連は最も強い。

表 4　各変数の相互関係（相関係数、n=47）

		自殺率	高齢化率	完全失業率	県民所得
自殺率	偏差積和S_{xy}	237.836	92.811	8.323	-11.461
	共分散s_{xy}	5.060	1.975	0.177	-0.244
	相関係数r_{xy}	1	0.321	0.132	-0.216
高齢化率	偏差積和S_{xy}	92.811	352.220	77.010	-26.445
	共分散s_{xy}	1.975	7.494	1.639	-0.563
	相関係数r_{xy}	0.321	1	-0.362	-0.410
完全失業率	偏差積和S_{xy}	8.323	77.010	16.837	-2.507
	共分散s_{xy}	0.177	1.639	0.358	-0.053
	相関係数r_{xy}	0.132	-0.362	1	-0.178
県民所得	偏差積和S_{xy}	-11.461	-26.445	-2.507	11.810
	共分散s_{xy}	-0.244	-0.563	-0.053	0.251
	相関係数r_{xy}	-0.216	-0.410	-0.178	1

（3）高齢化率から自殺率を説明する回帰分析

1）回帰係数の表

以下、高齢化率から自殺率を説明する回帰分析を例に解説を行うことにしよう。統計ソフトによって表の構成に多少の違いはあるが、回帰分析をおこなった場合は回帰係数の表、平方和に関する表が示される。このうち、回帰係数の表が次の表 5 である。表には網掛け部分があるが、これは推測統計に関わる結果であり、現時点では触れない（5 節で解説する)。

表 5　回帰係数の表（高齢化率→自殺率）

	非標準化係数	標準誤差	標準化係数	t	有意確率
高齢化率	0.264	0.116	0.321	2.271	0.028
［定数］	11.391	3.297		3.455	0.001
決定係数	0.103				

上の表のうち「非標準化係数」の列に示されているのが、標準化されていないデータを用いた場合の分析結果である。「高齢化率」の行に示されているのが回帰係数（b_1）の値、［定数］の行に示されているのが切片（b_0）の値だ。したがって、標準化されていないデータを用いて高齢化率から自殺率を説明する回帰直線の式は、表 5 から次のようになることがわかる。

$$Y = 11.391 + 0.264X$$

この結果は表 3 と表 4 からも導き出せる。回帰係数と切片は次の式で求められた。

$$\begin{cases} b_1 = \dfrac{S_{xy}}{S_{xx}} \\ b_0 = \bar{y} - \dfrac{S_{xy}}{S_{xx}}\bar{x} \end{cases}$$

ここから

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{92.811}{352.220} = 0.264$$

$$b_0 = \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x}$$

$$= 18.845 - \frac{92.811}{352.220} \times 28.285 = 11.391$$

となり、回帰直線は

$$Y = 11.391 + 0.264X$$

と導き出せるのである。

さて、表 5 の「標準化係数」の列に示されているのは、標準化したデータを用いた場合の分析結果、すなわち標準回帰係数だ。標準化したデータを用いて高齢化率から自殺率を説明する回帰直線の式は

$$Y = 0.321X$$

である（切片がないことに注意）。非標準化の回帰係数を見ても標準回帰係数を見ても、高齢化率は自殺率に正の影響力を持っていることがわかる。つまり、高齢化率が高くなるほど自殺率も高くなる傾向があるといえる。

ところで、標準回帰係数を説明する際に「x と y の標準偏差が同じ場合、回帰係数は両者の相関係数と一致する」と述べた。標準回帰係数は、今見た通り 0.321 だ。他方、表 4 の相関係数を見ると、こちらも 0.321 である。標準回帰係数と相関係数は確かに同じ値になっている。

標準回帰係数をbとするとき、非標準化回帰係数b_1と 2 つの標準偏差s_x、s_yを用いた次の式が成り立つ。

$$b = b_1 \frac{s_x}{s_y}$$

それゆえ、仮に分析結果で非標準回帰係数しか提示されていなかったとしても、x と y の標準偏差の情報があれば自分で標準回帰係数を求めることができる。例えば、表 5 の非標準化係数の値と表 3 の標準偏差の値を用いて計算すると、標準回帰係数は、

$$b = 0.264 \times \frac{2.738}{2.250} = 0.321$$

となる。これは表 5 の標準化係数の値と一致していることが確認できるだろう。

2）平方和の表

統計ソフトで回帰分析を行う場合、通常、表 6 のような平方和の表も表示される（こちらの表にも、推測統計に関わる網掛け部分があるが、これも 5 節で解説する）。

表には、「回帰」と「残差」、「合計」それぞれの平方和が示されているが、これは 3.3 で述べた回帰平方和、残差平方和、全平方和のことである。すなわち、

$$S_R = \sum (\hat{y}_i - \bar{y})^2 = 24.456$$
$$S_e = \sum (y_i - \hat{y}_i)^2 = 213.380$$
$$S_{yy} = \sum (y_i - \bar{y})^2 = 237.836$$

である。全平方和は 237.836 となっているが、これは表 3 の自殺率の偏差平方和に等しい。これをまず確認してほしい。

表 6　平方和の表（高齢化率→自殺率）

	平方和	自由度	平均平方	F	有意確率
回帰	24.456	1	24.456	5.158	0.028
残差	213.380	45	4.742		
合計	237.836	46			

（回帰平方和）÷（全平方和）が決定係数である。これはデータの全変動を回帰モデルがどの程度説明できるかを表したものだ。表 6 をもとに計算すると、決定係数は

$$R^2 = \frac{S_R}{S_{yy}} = \frac{24.456}{237.836} = 0.103$$

となる。これは確かに表 5 で示されている決定係数の数値に一致している。表 3 を見ると、決定係数は高齢化率と自殺率の相関係数の 2 乗と一致していることもわかる。

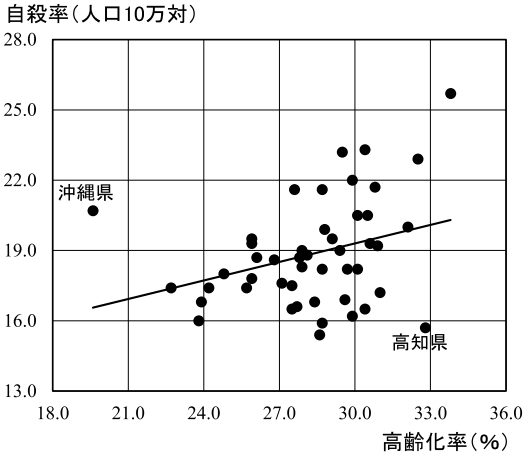
$$0.321^2 = 0.103$$

3）散布図と回帰直線

さて、高齢化率と自殺率の関連は、散布図では次のようになっている（図 8）。

この散布図でみると、沖縄県（19.6，20.7）や高知県（32.8，15.7）などが外れ値になっており、回帰直線がこれに引きずられている様子が分かる。実際、これらを除外した場合、回帰係数は 0.513、決定係数は 0.263 になる。決定係数が元の 0.103 から 0.263 に代わるということは、これらの除外によって、自殺率を 10％しか説明できなかったこの回帰モデルが、26％まで説明できるようになったということを意味する。

図 8　散布図と回帰直線（高齢化率→自殺率）



3.5　課題

（1）問題

上のデータを用い、県民所得から自殺率を予測するような回帰分析を統計ソフトで行ったところ、表 7 および表 8 のような結果が出力された。これらの表を見て以下の問いに答えよ。

a）標準化されていない回帰係数等をもとに、回帰直線の式を書け。
b）標準化されている回帰係数等をもとに、回帰直線の式を書け。
c）平方和の分析の表から決定係数を算出せよ。
d）県民所得と自殺率の相関係数、この回帰分析の決定係数、この回帰分析の標準化された回帰係数の関係を述べよ。

表 7　回帰係数の表（県民所得→自殺率）

	非標準化係数	標準誤差	標準化係数	t	有意確率
県民所得	-0.970	0.653	-0.216	-1.486	0.144
［定数］	21.634	1.905		11.354	0.000
決定係数	0.047				

表 8　平方和の表（県民所得→自殺率）

	平方和	自由度	平均平方	F	有意確率
回帰	11.122	1	11.122	2.208	0.144
残差	226.714	45	5.038		
合計	237.836	46			

（2）解答

a）

$$Y = 21.634 - 0.970X$$

b）

$$Y = -0.216X$$

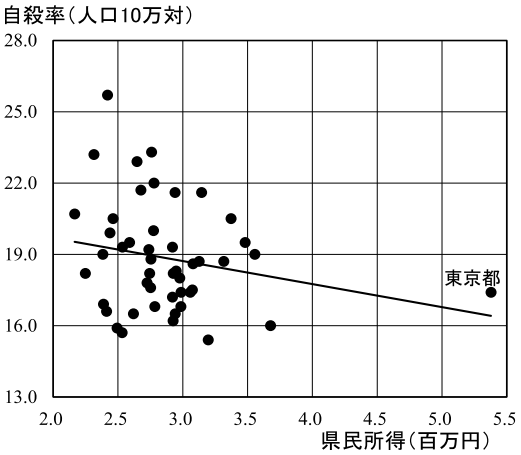
c）

$$\frac{11.122}{237.836} = 0.047$$

d）県民所得と自殺率の相関係数は‐0.216 であり、これが標準回帰係数と一致する。また、相関係数を 2 乗した（-0.216）^{2}＝0.047 が、決定係数と一致する。

参考までにこの問題についての散布図と回帰直線を図 9 に示しておく。

図 9　散布図と回帰直線（県民所得→自殺率）



4　回帰分析の図形による理解

4.1 変数ベクトルと平方和・相関係数

（1）変数ベクトル

回帰分析はベクトルを用いて図形的にとらえることでイメージがわきやすくなる。ここでは、相関係数や回帰係数、決定係数、平方和の式について､ベクトルを用いて図形的な解釈を行う｡まず、平均偏差得点をもとにしたベクトルから話を進める。平均偏差得点とはそれぞれの素点から平均を引いた得点のことである（この節では偏差得点と略すこともある)。

表 9 には、1 氏～5 氏の 5 人の年齢と収入についての素点と平均偏差得点、ならびに基本的な統計量が示されている。

表 9　素点と平均偏差得点

	素点		平均偏差得点	
	年齢	収入	年齢	収入
1氏	28	550	-11.6	-68.0
2氏	35	560	-4.6	-58.0
3氏	36	650	-3.6	32.0
4氏	49	630	9.4	12.0
5氏	50	700	10.4	82.0
n	5	5	5	5
平均	39.6	618.0	0.0	0.0
分散	73.0	3176.0	73.0	3176.0
標準偏差	8.5	56.4	8.5	56.4
偏差平方和	365.2	15880.0	365.2	15880.0
相関係数	0.79		0.79	
偏差積和	1906		1906	
共分散	381.2		381.2	

この平均偏差得点をベクトルとして表現できることに注目しよう。年齢も収入も、それぞれ 5 次元（ケース数）のベクトルとして表現できる。たとえば年齢ベクトルと収入ベクトルは次のようになる（ベクトルはこのように太字で表現される)。

$$\mathbf{x}=\begin{bmatrix}-11.6\\-4.6\\-3.6\\9.4\\10.4\end{bmatrix}\qquad \mathbf{y}=\begin{bmatrix}-68.0\\-58.0\\-32.0\\12.0\\82.0\end{bmatrix}$$

これらの変数を表すベクトルを（偏差得点の）変数ベクトルと呼ぶことにする。変数ベクトルは n 次元（ケース数次元）のベクトルである。

（2）変数ベクトルと平方和

変数ベクトルの長さ（大きさ）は、その変数の偏差平方和（変動）の平方根になる。

$$\|\mathbf{x}\|=\sqrt{S_{xx}}$$

というのは、ベクトルの長さは、

$$\|\mathbf{x}\|=\sqrt{x_1^2+x_2^2+\cdots+x_n^2}$$

であり、平均偏差得点のときには偏差平方和について次のことが成り立つからだ。

$$\sqrt{S_{xx}}=\sqrt{\sum\left(x_i-\bar{x}\right)^2}=\sqrt{\sum\left(x_i-0\right)^2}$$
$$=\sqrt{\sum x_i^{\,2}}=\sqrt{x_1^2+x_2^2+\cdots+x_n^2}$$

（3）変数ベクトルと相関係数

変数ベクトルは相関係数にも関係している。2 つの偏差得点化された変数ベクトルの余弦は、実は相関係数に他ならない。どうしてそんなことに

なるのかを説明しよう。

相関係数は次のようなものだった。

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

この分子である偏差積和は偏差得点を用いる場合は次のようになる。

$$\begin{aligned} S_{xy} &= \sum (x_i - \bar{x})(y_i - \bar{y}) \\ &= \sum (x_i - 0)(y_i - 0) = \sum x_i y_i \end{aligned}$$

これは、ベクトルの内積の定義と同じである。

$$(\mathbf{x}, \mathbf{y}) = \sum x_i y_i$$

ここで、相関をベクトルの内積と大きさで表現すると次のようになる。

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}} = \frac{(\mathbf{x}, \mathbf{y})}{\|\mathbf{x}\|\|\mathbf{y}\|}$$

この右辺は、高校で習ったベクトル間の角度 θ に関する次の公式の右辺と同一である。

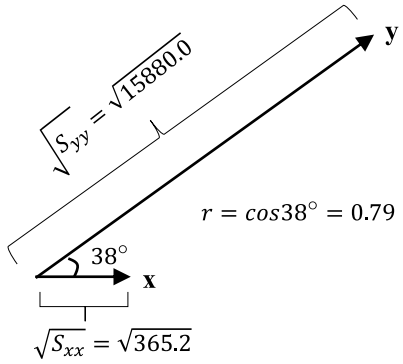
$$\cos\theta = \frac{(\mathbf{x}, \mathbf{y})}{\|\mathbf{x}\|\|\mathbf{y}\|}$$

したがって、２つの変数ベクトルが作る角度の余弦は相関係数に等しいといえる。

さて、以上のことを図形的に考えてみよう。2 次元ベクトルや 3 次元ベクトルなら図に描けるが、それを超える次元のベクトルは図に描けないと思うかもしれない。しかし、何次元ベクトルでもベクトルが 2 本までだと、それらの関係は平面に描けるし、3 本までだと空間に描ける。

このことを念頭に置き、先の年齢ベクトルと収入ベクトルを図示してみよう。表 9 から、年齢と収入の偏差平方和はそれぞれ 365.2 と 15880.0 であること、両者の相関は 0.79 であることがわかる。これらをもとに 2 つの変数ベクトルの関係を描くと図 10 のようになる。原点（0,0,0,0,0）は素点でいうと年齢の平均、収入の平均をともに表す位置であり、そこを起点に 2 つの変数ベクトルが伸びていくというイメージをもってほしい。どのような偏差得点をもとにした 2 つの変数ベクトルも、このような形で図示できるのである。

図 10　2 つの変数ベクトルと相関係数



相関にかかわるさまざまな問題を感覚的に理解するためには「相関＝ベクトル間の角度の余弦」というとらえ方はとても有効だ。2 つのベクトルが作る角度と相関の値との関係を表にまとめておく（表 10）。

表を見るとわかるように、相関が 1 のとき 2 つの変数ベクトルは同じ方向を向いており、相関が −1 のとき逆方向を向いている。相関が 0 であるとき、2 つの変数ベクトルは直交する。

表 10　変数の相関と変数ベクトル間の角度

相関	角度	相関	角度	相関	角度
1.00	0.0	0.30	72.5	-0.40	113.6
0.95	18.2	0.25	75.5	-0.45	116.7
0.90	25.8	0.20	78.5	-0.50	120.0
0.85	31.8	0.15	81.4	-0.55	123.4
0.80	36.9	0.10	84.3	-0.60	126.9
0.75	41.4	0.05	87.1	-0.65	130.5
0.70	45.6	0.00	90.0	-0.70	134.4
0.65	49.5	-0.05	92.9	-0.75	138.6
0.60	53.1	-0.10	95.7	-0.80	143.1
0.55	56.6	-0.15	98.6	-0.85	148.2
0.50	60.0	-0.20	101.5	-0.90	154.2
0.45	63.3	-0.25	104.5	-0.95	161.8
0.40	66.4	-0.30	107.5	-1.00	180.0
0.35	69.5	-0.35	110.5		

（4）変数ベクトルと回帰係数

変数ベクトルから回帰係数の図形的解釈もできる。回帰の予測式はこれまで次のように表されてきた。

$$\hat{y}_i = b_0 + b_1 x_i$$

得点を平均偏差得点に変換すると、次のように切片はなくなり、回帰係数だけの式になる。

$$\hat{y}_i = \left(\bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x} \right) + \frac{S_{xy}}{S_{xx}} x_i$$
$$= \left(0 - \frac{S_{xy}}{S_{xx}} 0 \right) + \frac{S_{xy}}{S_{xx}} x_i = \frac{S_{xy}}{S_{xx}} x_i$$

そこで、ここからは回帰係数をb_1ではなくbと表現して話を進める。

$$\hat{y}_i = b x_i$$

$$b = \frac{S_{xy}}{S_{xx}}$$

素点をもとにしても平均偏差得点をもとにしても、この回帰係数の値は変わらない。平均偏差得点にして変わるのは切片が 0 になるというところだけである。表 9 のデータからbを計算すると次のようになる。

$$b = \frac{S_{xy}}{S_{xx}} = \frac{1906}{365.2} = 5.22$$

さて、すべてのx_iとそれに対応する$\hat{y}_i$についての回帰の予測式をベクトルで表すと次のようになる（ベクトルは太字で表す）。

$$\hat{\mathbf{y}} = b\mathbf{x}$$
$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = b \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

また、実際の値と予測値の差である残差ベクトルは次のようになる。

$$e = \mathbf{y} - \hat{\mathbf{y}}$$
$$\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}$$

そして、そして回帰分析では次の事柄が成立している。

$$\mathbf{y} = b\mathbf{x} + \mathbf{e} = \hat{\mathbf{y}} + \mathbf{e}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = b \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

この最後の式を、これまで見てきた年齢と収入の偏差得点による変数ベクトルで表すと次のようになる。

$$\mathbf{y} = b\mathbf{x} + \mathbf{e} = \hat{\mathbf{y}} + \mathbf{e}$$

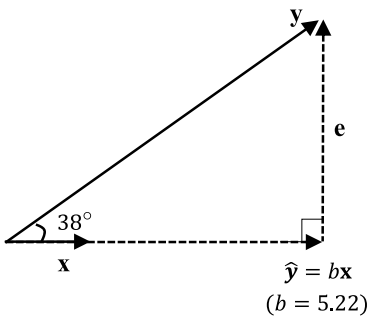
$$\begin{bmatrix} -68.0 \\ -58.0 \\ -32.0 \\ 12.0 \\ 82.0 \end{bmatrix} = 5.22 \begin{bmatrix} -11.6 \\ -4.6 \\ -3.6 \\ 9.4 \\ 10.4 \end{bmatrix} + \begin{bmatrix} -7.5 \\ -34.0 \\ 50.8 \\ -37.1 \\ 27.7 \end{bmatrix}$$

$$= \begin{bmatrix} -60.5 \\ -24.0 \\ -18.8 \\ 49.1 \\ 54.3 \end{bmatrix} + \begin{bmatrix} -7.5 \\ -34.0 \\ 50.8 \\ -37.1 \\ 27.7 \end{bmatrix}$$

このベクトルの演算で示されている関係を図で表すと図 11 のようになる。この図において **x** は年齢ベクトル、**y** は実際の収入のベクトル、$\hat{\mathbf{y}}$ は収入の予測値のベクトルである。$\hat{\mathbf{y}}$ は原点を起点とし、**y** の先端から **x** ベクトル方向に垂線を下ろした足を終点とするベクトルだ。回帰係数 b とは、$\hat{\mathbf{y}}$ ベクトルの長さが **x** ベクトルの何倍かを示したものである。この図から、次のベクトルの演算が成り立っていることがわかる。

$$\mathbf{y} = b\mathbf{x} + \mathbf{e} \qquad \mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}$$

図 11　回帰係数の図形表現

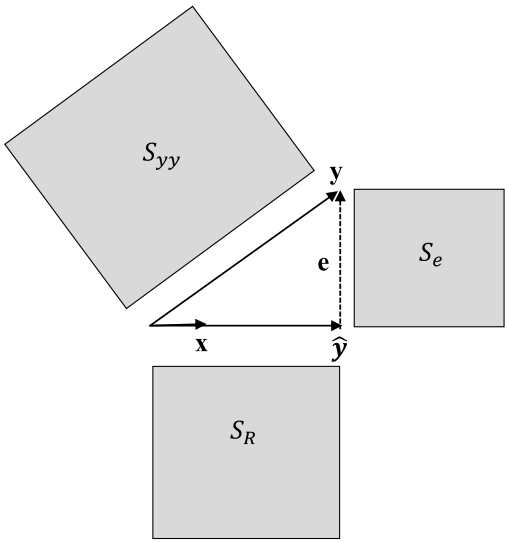


4.2 変数ベクトル・平方和・決定係数

（1）平方和と面積

変数 y の偏差平方和（変動）S_{yy} は、変数ベクトル **y** の長さの 2 乗、$\|\mathbf{y}\|^2$ になる。それは図 12 の面積 S_{yy} として把握できる。

図 12　回帰分析の平方和の図形表現



同様に、残差平方和（残差変動）S_eや回帰平方和（回帰変動）S_Rも面積として把握できる。ここで、ピタゴラスの定理より$S_{yy} = S_R + S_e$が成立しているため、全平方和＝回帰平方和＋残差平方和という等式が成り立つことがすぐにわかる。y の変動S_{yy}は、x によって説明される変動S_Rと説明されない変動S_eにきれいに分解されるのである。

先の例を使ってこのことをもう少し検討しよう。次のような変数ベクトルがあった。

$$\mathbf{y} = \begin{bmatrix} -68.0 \\ -58.0 \\ -32.0 \\ 12.0 \\ 82.0 \end{bmatrix} \quad \hat{\mathbf{y}} = \begin{bmatrix} -60.5 \\ -24.0 \\ -18.8 \\ 49.1 \\ 54.3 \end{bmatrix} \quad \mathbf{e} = \begin{bmatrix} -7.5 \\ -34.0 \\ 50.8 \\ -37.1 \\ 27.7 \end{bmatrix}$$

これを用いると、それぞれに対応する面積は次のようになって、全平方和＝回帰平方和＋残差平方和に実際になっていることがわかる。

$$S_{yy} = \|\mathbf{y}\|^2 = \begin{Bmatrix} (-68.0)^2 + (-58.0)^2 \\ +(-32.0)^2 + 12.0^2 \\ +82.0^2 \end{Bmatrix} = 15880.0$$

$$S_R = \|\hat{\mathbf{y}}\|^2 = \begin{Bmatrix} (-60.5)^2 + (-24.0)^2 \\ +(-18.8)^2 + 49.1^2 \\ +54.3^2 \end{Bmatrix} = 9947.5$$

$$S_e = \|\mathbf{e}\|^2 = \begin{Bmatrix} (-7.5)^2 + (-34.0)^2 \\ 50.8^2 + (-37.1)^2 \\ +27.7^2 \end{Bmatrix} = 5932.5$$

そして、

$$S_R + S_e = 9947.5 + 5932.5 = 15880.0 = S_{yy}$$

となっていることがわかる。

（2）決定係数

先の面積の図（図 12）からは、S_R / S_{yy} は y の変動の内で x で説明される変動の割合ということがすぐにわかるだろう。これが決定係数R^2である。

$$R^2 = \frac{S_R}{S_{yy}}$$

図をもう少し詳しく見てみると、決定係数について次の式が成り立つことがわかる。

$$R^2 = \frac{S_R}{S_{yy}} = \left(\frac{\|\hat{\mathbf{y}}\|}{\|\mathbf{y}\|} \right)^2 = \cos^2 \theta = r^2$$

すなわち、決定係数は相関係数の 2 乗である。

年齢と収入の値を使って決定係数を計算すると、

$$R^2 = \frac{S_R}{S_{yy}} = \frac{9947.5}{15880.0} = 0.626$$

となっており、確かに相関係数（.79）の 2 乗になっている。

決定係数は「y の変動（平方和）の内で x で説明される変動の割合」と述べたが、変動ではなく分散から考えてもよい。というのは次の式が成り立つからだ。

$$R^2 = \frac{S_R}{S_{yy}} = \frac{n s_{\hat{y}}^2}{n s_y^2} = \frac{s_{\hat{y}}^2}{s_y^2}$$

決定係数は「y の分散の内で x で説明される分散の割合」でもあるのだ。

4.3 標準得点の変数ベクトル

（1）偏差得点ベクトルと標準得点ベクトル

これまでは平均偏差得点をもとにした変数ベクトルについて見てきたが、ここからは標準得点をもとに考えよう。標準得点とは、変数の平均が 0、分散が 1 になるように変換した得点だ。標準偏差が 1 のとき分散も 1 だから、分散が 1 になるように調整したものと考えてもいい。表 11 は 1 氏～5 氏の 5 人の年齢と収入についての素点と標準得点、ならびに基本的な統計量を示したものだ。

表 11　素点と標準得点

	素点		標準得点	
	年齢	収入	年齢	収入
1氏	28	550	-1.36	-1.21
2氏	35	560	-0.54	-1.03
3氏	36	650	-0.42	0.57
4氏	49	630	1.10	0.21
5氏	50	700	1.22	1.46
n	5	5	5	5
平均	39.6	618.0	0.00	0.00
分散	73.0	3176.0	1.00	1.00
標準偏差	8.5	56.4	1.00	1.00
偏差平方和	365.2	15880.0	5.00	5.00
相関係数	0.79		0.79	
偏差積和	1906		3.96	
共分散	381.2		0.79	

標準得点化された年齢と収入のベクトルは次のようになる。

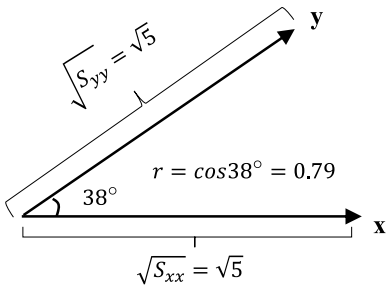
$$\mathbf{x}=\begin{bmatrix}-1.36\\-0.54\\-0.42\\1.10\\1.22\end{bmatrix}\qquad \mathbf{y}=\begin{bmatrix}-1.21\\-1.03\\0.57\\0.21\\1.46\end{bmatrix}$$

標準得点化をもとにしたベクトルについては、上の偏差得点のベクトルについて述べたことがすべて成り立つ。すなわち、各変数ベクトルの大きさは変数の偏差平方和（変動）の正の平方根である。また、変数ベクトルの作る角の余弦は相関係数に等しい。

標準得点についてさらにいえることは、両変数の標準偏差が 1 になっているため、各変数ベクトルの大きさが等しくなり、その大きさはケース数の平方根になるということである。

標準得点化された年齢ベクトル **x** と収入ベクトル **y** を図示すると、下のように等しい長さのベクトルになる（図 13）。ここでも、$\cos\theta$ は x と y との相関係数の値になる。

図 13　標準化された変数ベクトルと相関係数



（2）標準得点による回帰分析の図形表現

年齢と所得を標準得点に変換した場合、回帰の予測式は次のようになる（回帰係数が相関係数と一致していることにも注意）。

$$\hat{y}_i = 0.79x_i$$

このとき、すべての x_i とそれに対応する $\hat{y}_i$ についての回帰の予測式をベクトルで表すと次のようになる。

$$\hat{\mathbf{y}} = 0.79\mathbf{x}$$

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \hat{y}_4 \\ \hat{y}_5 \end{bmatrix} = 0.79 \begin{bmatrix} -1.36 \\ -0.54 \\ -0.42 \\ 1.10 \\ 1.22 \end{bmatrix}$$

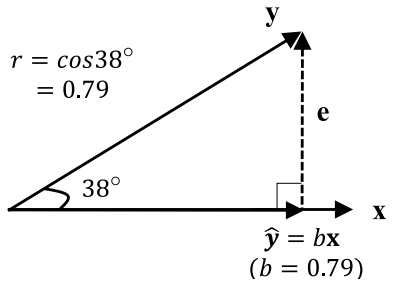
そして回帰式をベクトルで表すと次のようになる。

$$\mathbf{y} = b\mathbf{x} + \mathbf{e} = \hat{\mathbf{y}} + \mathbf{e}$$

$$\begin{bmatrix} -1.21 \\ -1.03 \\ 0.57 \\ 0.21 \\ 1.46 \end{bmatrix} = 0.79 \begin{bmatrix} -1.36 \\ -0.54 \\ -0.42 \\ 1.10 \\ 1.22 \end{bmatrix} + \begin{bmatrix} -0.13 \\ -0.60 \\ 0.90 \\ -0.66 \\ 0.49 \end{bmatrix}$$

$$= \begin{bmatrix} -1.07 \\ -0.43 \\ -0.33 \\ 0.87 \\ 0.96 \end{bmatrix} + \begin{bmatrix} -0.13 \\ -0.60 \\ 0.90 \\ -0.66 \\ 0.49 \end{bmatrix}$$

図 14　標準回帰係数の図形表現



このベクトルの演算で示されている関係を図で表すと図 14 のようになる。この図においても、図 11 と同様、**x** は年齢ベクトル、**y** は実際の収入のベクトル、$\hat{\mathbf{y}}$ は収入の予測値のベクトルである。$\hat{\mathbf{y}}$ は原点を起点とし、**y** の先端から **x** ベクトル方向に垂線を下ろした足を終点とするベクトルだ。

（3）標準回帰係数

標準得点を用いたときの回帰係数を標準回帰係数という。図 14 にある b は標準回帰係数だ。

$$b = \frac{\|\hat{\mathbf{y}}\|}{\|\mathbf{x}\|}$$

x ベクトル方向にある $\hat{\mathbf{y}}$ ベクトルの長さを「**x** ベクトルの長さの何割なのか」という観点から示したものが標準回帰係数 b である。**y** ベクトルが垂直に近づき θ が大きくなれば、$\hat{\mathbf{y}}$ ベクトルは短くなり b は 0 に近づく。**y** ベクトルが **x** ベクトルに近づき θ が小さくなれば、$\hat{\mathbf{y}}$ ベクトルの長さは長くなり、b は 1 に近づいていく。図からわかるように、b の値は 1 を超えない。範囲は−1 から＋1 である（標準化されていない回帰係数にはそのような制限はない）。

θ の余弦は x と y の相関係数だが、標準得点を用いた変数ベクトルの場合、**x** ベクトルと **y** ベクトルの大きさが等しくなるので次の式が成り立つ（図 14）。

$$b = \frac{\|\hat{\mathbf{y}}\|}{\|\mathbf{x}\|} = \frac{\|\hat{\mathbf{y}}\|}{\|\mathbf{y}\|} = \cos\theta = r$$

したがって、標準回帰係数は相関係数と等しくなる（図 11 を見るとわかるが、標準化されていない回帰係数ではそうはならない）。

（4）平方和と決定係数

標準得点をもとにした変数ベクトルで考えても、

平方和や決定係数のあり方は、偏差得点のベクトルの場合と同様である。

$$\mathbf{y}=\begin{bmatrix}-1.21\\-1.03\\0.57\\0.21\\1.46\end{bmatrix}\quad \hat{\mathbf{y}}=\begin{bmatrix}-1.07\\-0.43\\-0.33\\0.87\\0.96\end{bmatrix}\quad \mathbf{e}=\begin{bmatrix}-0.13\\-0.60\\0.90\\-0.66\\0.49\end{bmatrix}$$

これらの標準得点の変数ベクトルから、それぞれに対応する面積は次のようになって、「全平方和＝回帰平方和＋残差平方和」が実際に成立していることがわかる。

$$S_{yy}=\|\mathbf{y}\|^2=\left\{\begin{matrix}(-1.21)^2+(-1.03)^2\\+0.57^2+0.21^2\\+1.46^2\end{matrix}\right\}=5$$

$$S_R=\|\hat{\mathbf{y}}\|^2=\left\{\begin{matrix}(-1.07)^2+(-0.43)^2\\+(-0.33)^2+0.87^2\\+0.96^2\end{matrix}\right\}=3.13$$

$$S_e=\|\mathbf{e}\|^2=\left\{\begin{matrix}(-0.13)^2+(-0.60)^2\\+0.90^2+(-0.66)^2\\+0.49^2\end{matrix}\right\}=1.87$$

$$S_R+S_e=3.13+1.87=5=S_{yy}$$

また、決定係数は、

$$R^2=\frac{S_R}{S_{yy}}=\frac{3.13}{5}=0.626$$

であり、偏差得点のベクトルの場合に等しい。

4.4　課題

（1）問題

a) x で y を説明する回帰分析を行うとする。今、標準得点化された変数ベクトル **x** と **y** の作る角度が 30 度だとすると、そのときの x と y の相関係数はどうなるか、標準化係数を用いた回帰直線の式はどうなるか、また、この回帰分析における決定係数はどうなるか。

b) サイズ 100 の標本の回帰分析結果で、標準回帰係数が 0.5 となっていた。この回帰分析を標準得点化された 2 つの変数ベクトル **x**,**y** を使って作図せよ。2 つのベクトルとその長さ、ベクトル間の角度、ならびに予測値のベクトル $\hat{\mathbf{y}}$ を書いておくこと。

（2）解答

a)

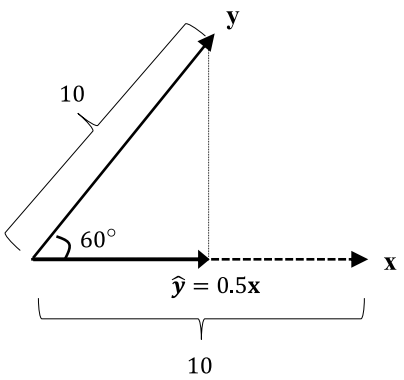
$$\cos 30^\circ=\frac{\sqrt{3}}{2}=0.866$$

$$r=b=0.866$$

$$Y=0.866X$$

$$R^2=(0.866)^2=0.750$$

b)



5　回帰分析の推測統計的理解

5.1 推測統計の基本

（1）母集団・標本・確率変数

1）確率変数とその分布

変数がとりうる値のそれぞれについて確率が決まっていることがある。そういう変数のことを確率変数という。この確率変数について、とりうる値とその確率の対応関係を示したものが確率分布である。すべての値に対する確率の計は 1 になる。表 12 は「さいころの目の数」という確率変数 X についての確率分布を示す表である。

表 12　さいころの目の確率分布表

確率変数X（さいころの目）

値(x)	1	2	3	4	5	6	計
確率(p)	1/6	1/6	1/6	1/6	1/6	1/6	1

確率変数には値が離散的な数値になるものと連続的な数値になるものがある。サイコロの目は離散型の確率変数であり、身長や体重などは連続型の確率変数である。ここでは連続型の確率変数に焦点を置き、どのように母平均が推定されるのかを説明する。本セミナーは回帰分析を取り扱うものだが、まずは母平均の推定の原理を学び、確率変数や確率分布の基礎知識と、検定や推定の原理を押さえておこうというわけである。

2）母集団と確率変数

ある大学の男子の身長という母集団があるとする。この母集団について 160cm 台は○人、170cm 台は△人などということがわかっているとしよう。これを全体での比率でとらえると、160cm 台は●割、170cm 台は▲割などということができる。このように値の比率からとらえた身長の分布は確率分布と考えることができる。というのは、160cm 台の比率は、その母集団から 1 人選ぶときにその人が 160cm 台である確率と同じだからだ。したがって、この場合、身長は確率変数なのである。ここでは身長を確率変数 x としよう。

1 つの確率変数の平均のことをその変数の期待値という。この母集団での身長の平均は身長の期待値というわけだ。この値が 170 のとき、身長の期待値は 170 といい、E(x)=170 と表す。母集団における期待値を母平均と呼び、μ で表す。μ ＝E(x)である。母集団の身長の分布が定まっていれば、その期待値は特定の「決まった値」になる。

またこの母集団での身長の分散が 100 のとき、V(x)=100 と書く。母集団での分散は母分散とよばれ、σ^2 で表す。σ^2＝V(x)である。母集団の身長分布が定まっていれば、その分散は特定の「決まった値」になる。

分布の期待値や分散は、分布が定まっている限り決まった値になる。このことはとても重要である。母平均、母分散、母標準偏差などのことを母数というが、これらは決まった値を持つものだ。

3）標本と確率変数

さて、この母集団からサイズ 10 の標本を抽出することを考えよう。実際に標本を抽出するのではなく、抽出することを考えるのだ。その場合抽出される標本は実際の値からなる集合ではなく、次のような確率変数の集合となる。

$$\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\}$$

推測統計では標本は確率変数の集合である。このことは必ず押さえておかなければならない。

（2）標本統計量と母数の推定

1）標本平均・標本分散・不偏分散

■標本平均

さて、上の標本における平均は次の式で算出される。

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

これは 1 つの確率変数の期待値という意味でなく、複数の確率変数の平均であることに注意しよう。上の身長の標本については、

$$\bar{x} = \sum_{i=1}^{10} x_i = \frac{1}{10}\left(x_1 + x_2 + \cdots + x_{10}\right)$$

ここで、$\bar{x}$ のことを（身長の）標本平均という。標本平均は x_1〜x_n という確率変数の平均だから、それ自体確率変数である。

■標本分散

また、標本における分散のことを標本分散 s^2 といい、次の式で表される。

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

これもまた確率変数である。

■不偏分散

不偏分散は次のものだ。

$$u^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

これもまた確率変数である。

標本平均、標本分散、不偏分散など、標本から算出される統計量のことを標本統計量という。推測統計的に考える場合、これらが決まった値をもつものではなく、確率変数であることに注意が必要である。

2）母数の推定

標本統計量は母平均や母分散などの母数の推定に利用される。母数の推定に用いられる標本統計量を推定量という。標本平均は母平均の推定量であり、標本分散や不偏分散は母分散の推定量である。推定量のうち、その期待値が母数に等しくなるものを不偏推定量という。

標本平均は母平均の不偏推定量であり、不偏分散は母分散の不偏推定量である。標本分散は母分散の推定量だが不偏推定量ではない。すなわち、

$$E\left(\bar{x}\right) = \mu \qquad E\left(u^2\right) = \sigma^2 \qquad E\left(s^2\right) \neq \sigma^2$$

現実の標本の値を用いて算出した推定量の値を推定値という。

3）標本統計量の期待値と分散

確率変数である標本平均、標本分散、不偏分散といった標本統計量はさまざまな値をとりうる。しかしながら、母集団の分布が決まっている場合には、標本統計量の確率分布は数学的に導ける。たとえば標本平均の場合、次のようなことがいえる。

「母集団における母平均が μ、母分散が σ^2 であるとき、そこから得られたサイズ n の標本の標本平均 $\bar{x}$ は、期待値が μ になり分散が σ^2/n になる」。

すなわち、

$$E\left(\bar{x}\right) = \mu \qquad V\left(\bar{x}\right) = \sigma^2/n$$

ここでは、標本統計量である標本平均$\bar{x}$が母平均μと同じ値を中心に分散σ^2/nの広がりで分布するとされているのである。

4）標準誤差

標本平均の分布の標準偏差$\sqrt{\sigma^2/n}$を標本平均$\bar{x}$の標準誤差という。「標準誤差」という用語は「推定量として用いる標本統計量（確率変数）の標準偏差」を指す。それは、標本平均だけでなくあらゆる推定量に用いられ、「○○の標準誤差」と表現される。この場合は「標本平均」の標準誤差なのだが、「標本比率」の標準誤差、「標本回帰係数」の標準誤差など、さまざまな「標準誤差」がある。

さて、標本平均の標準誤差は$\sqrt{\sigma^2/n}$だったのだが、ここには母数のσ^2が含まれる。したがって、標本のデータだけでは標準誤差は明らかにならない。そんなときには、母分散σ^2のかわりにその標本から得られた母分散の推定量である不偏分散u^2を使う。このようにして計算される$\sqrt{u^2/n}$も、標本平均$\bar{x}$の標準誤差である。

（3）標本統計量の確率分布

1）代表的な確率分布

確率分布の中には数学的に定式化されたものがある。推測統計で用いる代表的な連続型の確率分布は正規分布、χ^2分布、t 分布、F 分布である[1]。

Z を標準正規分布に従う確率変数とし、K_nを自由度 n のχ^2分布に従う確率変数とする。また、T_nを自由度 n の t 分布に従う確率変数とし、F_n^mを自由度 m,n の F 分布に従う確率変数とする。このとき次の式が成り立つ（右辺にある確率変数はすべて独立であるとする）。

$$K_1 = Z^2$$
$$K_n = Z_1^2 + Z_2^2 + \cdots + Z_n^2$$
$$T_n = \frac{Z}{\sqrt{K_n/n}}$$
$$F_n^m = \frac{K_m/m}{K_n/n}$$

上の式をよく見るとわかるはずだが、標準正規分布に従う確率変数の 2 乗は自由度 1 のχ^2分布に従う。また、自由度 n の t 分布に従う確率変数の 2 乗は自由度（1,n）の F 分布に従う。このように、これらの確率分布は相互に深い関係がある。確率分布間にあるこういった相互関係は次のように利用される。

2）標本統計量の確率分布

母集団 x が正規分布をしているときには標本平均$\bar{x}$も正規分布をする。すなわち、

$$x \sim N\left(\mu, \sigma^2\right) \quad \Rightarrow \quad \bar{x} \sim N\left(\mu, \sigma^2/n\right)$$

右側の標本平均のほうを標準化すると

$$x \sim N\left(\mu, \sigma^2\right) \quad \Rightarrow \quad \frac{\bar{x}-\mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

しかし、ここには母数のσ^2が含まれるので、標本のデータだけからはどのような分布になるかはわからない。そこで、母分散σ^2のかわりに標本統計量である不偏分散u^2を使う。しかし、こうすることによって、標本平均$\bar{x}$は標準正規分布にはした

[1] 小林（2018c）参照。

がわないことになる。それは、標準正規分布に形の似たt分布（自由度n−1）に従うことになるのである（ここに確率分布間の相互関係が利用されている。解説は次節に譲る）。

$$x \sim N\left(\mu, \sigma^2\right) \Rightarrow \frac{\bar{x}-\mu}{\sqrt{u^2/n}} \sim t_{n-1}$$

（4）検定と区間推定

1）検定

統計的仮説検定とは、標本データをもとに母集団についてあることがいえるかどうかをテストするものである。たとえば、ある大学の男子学生の身長の平均が170cmだと信じられているとき、標本調査をすると175cmあったとしよう。この標本での175cmという平均をもとに、「母集団である大学全体の男子学生の平均身長は本当は170cmではない」といえるかどうかを検討するのが、統計的仮説検定の一例だ。

検定には、これまで上で述べてきたことをまとめたような定理が用いられる。今の検定に関わる定理は次のものだ[2]。

■定理：母平均μ、母分散σ^2（未知）の正規分布に従う母集団からのサイズnの標本について、標本平均を$\bar{x}$、不偏分散をu^2とすれば、下の検定統計量Testは自由度n-1のt分布に従う。

$$Test = \frac{\bar{x}-\mu}{\sqrt{u^2/n}} \sim t_{n-1}$$

この調査での標本サイズが30、標本の不偏分散が100である場合、検定統計量は次のようになる。

$$\frac{175-170}{\sqrt{100/30}} = 2.7386$$

この3.536という値は自由度29のt分布においてはあまり出ない値である。母平均が170cmのとき、検定統計量がその値になる確率は5%以下だ（自由度29のt分布における両側確率5%の棄却域は±2.045の外側）。そこから有意水準5%で、母集団の平均身長は170cmではないと結論づけるのである。

2）区間推定

標本統計量から母数を推定する際、点として推定するだけでなく、どのような区間に母数があるのかを推定することがある。この区間を信頼区間という。上で見た平均についての検定統計量の式をもとにすると、母平均の95%の信頼区間は次のようになる。

$$\bar{x} - t_{n-1}(0.025)\frac{u}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1}(0.025)\frac{u}{\sqrt{n}}$$

この式は、確率変数としての標本平均を$\bar{x}$とするとき、母平均は95%の確率でこの範囲にあることを示すものである。

この$\bar{x}$に実際の標本で得られた標本平均の値を代入すると具体的な信頼区間が定まる。上の例の場合は、次のようになる。

[2] 小林（2019a）参照

$$175 - t_{30-1}(0.025)\frac{10}{\sqrt{30}} \le \mu$$
$$\le 175 + t_{30-1}(0.025)\frac{10}{\sqrt{30}}$$
$$175 - 2.045\frac{10}{\sqrt{30}} \le \mu$$
$$\le 175 + 2.045\frac{10}{\sqrt{30}}$$
$$\therefore 171.266 \le \mu \le 178.734$$

したがって、95%の信頼度で母平均は 171.266 と 178.734 の間にあるということになる。

ところで、ここで、確率という用語を用いず、信頼度という用語を用いていることに注意が必要だ。現実の標本の値を用いた母平均の区間の式は、当たっているかはずれているかのどちらかであり、確率の式とはいえないからである。信頼区間の「信頼度 95%」とは、「さまざまな現実の標本で信頼区間を算出した時、それらのうちの 95%は正しい式になる」という意味である。

図 15　信頼区間の意味

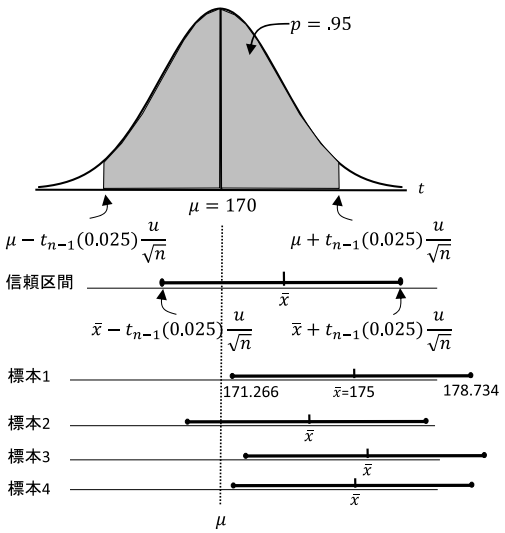


図 15 はこのあたりの事情を説明するためのものである。まず、下の標本 1 から標本 4 を見てほしい。今は上の確率分布は無視しよう。これに翻弄されてはいけない。

標本 1 は今回の例の標本であり、$\bar{x}$ は 175 である。別の標本で信頼区間を出せば標本 2 から標本 4 のように$\bar{x}$も異なる値になり、それぞれで信頼区間は異なることがわかるだろう。このように区間推定をたくさんしたとき、そのうち 95%のものに真の母平均が含まれるというのが信頼度の意味である。

さて図 15 の上の確率分布を見てみよう。これは、母平均＝170cm という帰無仮説の下での確率分布だ。これと下の標本の線分の関係についてよく見ると次のことがわかる。すなわち、ある現実の標本を用いた区間推定で 95%の信頼区間に 170cm が入るときには（標本 2）、その標本で「母平均＝170cm」という帰無仮説は棄却されない（有意確率 5%両側検定）。帰無仮説が棄却されるのはその信頼区間に 170cm が入っていない場合に限られる（標本 1,3,4）。このことを理解していると、標準誤差から信頼区間を想像できるだけでなく、帰無仮説の検定結果も想像できるようになる。

5.2 回帰分析の母集団と標本

（1）母集団のイメージ

上の説明で推測統計における母集団と標本の基本的な関係が理解できたと思う。ここからは回帰分析に焦点を置き、母集団と標本の関係について見ていく。

回帰分析ではいくつかの仮定が置かれている。それらの仮定を念頭に置き、母集団のイメージを示したものが図 16 である。色の濃いところに対象はたくさん分布し、色の薄いところにはあまり

分布していない。

濃いところが左から右に向けて直線的になっているところに注意してほしい。これが曲がった線ではなく直線であることが回帰分析の前提なのである。

また、X のどこでこの分布を縦に切り取っても、グラデーションの縦幅は同じであるということもわかる。グラデーションの縦幅が同じということは、どの x_i においても y_i の分布の分散は同じであることを意味している。たとえば、x_1 で切っても、x_2 で切ってもその分布の広がりは図 17 のようになっていて、どこでも同じなのである（図 16 の縦軸が図 17 の横軸になっていることに注意）。これもまた回帰分析の前提である。

図 16　回帰分析の母集団のイメージ

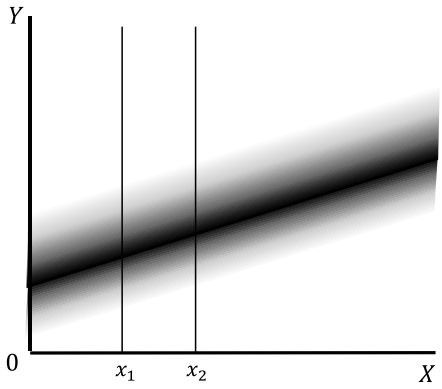
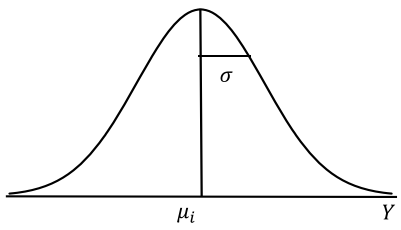


図 17　yi の確率分布



このような分布の具体的なイメージをつかむためには、成人男性の身長 X と体重 Y を思い描くといいかもしれない。どの身長の人も体重はある幅をもって分布し、その分布の幅はそうは変わらないと考えるのが自然だからだ。われわれはこういった母集団から標本を抽出し回帰分析を行うのである。

さて、以下では回帰分析における回帰式やそれに関連する式について解説する。ここでは次の母回帰式と標本回帰式がとりわけ重要である。

$$\begin{cases} y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \\ y_i = b_0 + b_1 x_i + e_i \end{cases}$$

（2）母集団の回帰式

1）母集団における y_i の期待値の式

母回帰式の解説の前に、「母集団における y_i の期待値の式」について解説しよう。この式を理解することが母回帰式の理解の前提となる。母集団における y_i の期待値の式は次のものである。

$$E(y_i) = \beta_0 + \beta_1 x_i$$
$$\mu_i = \beta_0 + \beta_1 x_i$$

y_i の期待値が $E(y_i)$ でありそれが μ_i という記号で表現されているので 2 つは同じ式である。

先のイメージで x_i それぞれにおいて y_i が確率分布をする様子を見たが、この各 y_i の確率分布の期待値が μ_i である。μ_i は期待値なので母集団の分布が定まっているならば、必ず決まった値になる。

$\mu_i = \beta_0 + \beta_1 x_i$ という 1 次式が母集団で成立しているというのが回帰分析の仮定である。β_0 , β_1 は母数であり、ともに決まった値である。 x_i は、横軸の x_i の場所を指定するだけのものなので確率変数ではない。したがって、$\mu_i = \beta_0 + \beta_1 x_i$ とい

う式は確率変数を含まない式である。以上をもとに、次に母回帰式について解説する。

2）母回帰式

母回帰式とは次の式である。

$$\boxed{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}$$

すでに述べたように、β_0 , β_1は決まった値をとり、 $\mu_i = \beta_0 + \beta_1 x_i$も決まった値をとる。 y_iはこのμ_iを中心に誤差ε_iを伴って現れる値である。

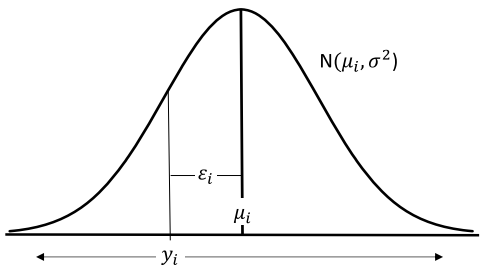
誤差ε_iは確率変数であり、その分布はどの i においても$N\left(0, \sigma^2\right)$であると仮定されている。そこから、$y_i$は$N\left(\mu_i, \sigma^2\right)$に従うことになる。

母回帰式におけるβ_1を母回帰係数（傾き）という。β_0 はここでは母回帰の切片ということにする。β_0 β_1を切片や傾きというのは、後で述べる回帰直線の式で切片や傾きになるからである。

3）誤差

誤差の仮定についてはすでに述べたが、y_iの分布とε_iの関係は図 18 のようになる。

図 18　yi の分布と εi



これが次の 3 つの式の示していることなのである。

$$\varepsilon_i = y_i - E\left(y_i\right) = y_i - \mu_i$$
$$\varepsilon_i \sim N\left(0, \sigma^2\right)$$
$$y_i \sim N\left(\mu_i, \sigma^2\right)$$

誤差ε_iの期待値は 0、分散はσ^2である。また、誤差の 2 乗ε_i^2の期待値は誤差の分散と等しくσ^2となる。

$$E\left(\varepsilon_i\right) = 0$$
$$E\left(\varepsilon_i^2\right) = \sigma^2$$
$$\left(\because V\left(\varepsilon_i\right) = \sigma^2 = E\left(\varepsilon_i - E\left(\varepsilon_i\right)\right)^2 = E\left(\varepsilon_i^2\right)\right)$$

誤差ε_iについては、異なる i についてのε_iは独立という仮定も設けられている。

$$Cov\left(\varepsilon_i, \varepsilon_j\right) = 0 \quad \left(i \neq j\right)$$

4）XY 軸グラフの母回帰直線の式

回帰分析を考えるに際しては、回帰直線をもとに考えていくとわかりやすい。ここでは母集団における母回帰直線を、大文字の X と Y を使って次の式で表現しよう。

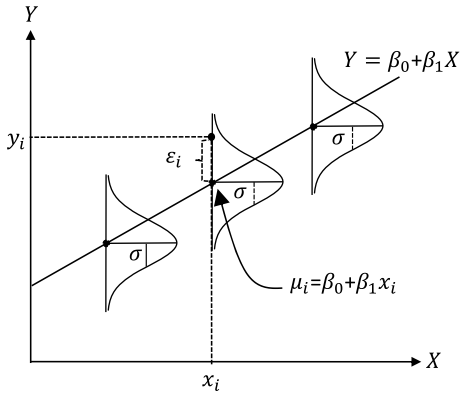
$$Y = \beta_0 + \beta_1 X$$

この式は、y_iの期待値$E\left(y_i\right)$すなわちμ_iについての式$\mu_i = \beta_0 + \beta_1 x_i$をもとにした式である。もとの式は、 x_iに対応する期待値の点を表すものだが、この母回帰直線の式はその点の集合を直線で表すものだ。母回帰直線は上で見た図 16 のイメージでの黒い部分の中心を通過する直線を意味する。

これまで母集団について述べたことをまとめると図 19 のようになる。この図をきちんと理解することが、回帰分析を推測統計的に考えるための

出発点になる。

図 19　母回帰直線と yi の分布



（3）標本の回帰式

1）標本回帰式

推測統計的に回帰分析を考える際の標本回帰式は次のものである。

$$y_i = b_0 + b_1 x_i + e_i$$

この式における x_i, y_i は母集団での回帰式の x_i, y_i と同じものである。それは、母集団にある x_i, y_i が抽出され、その x_i, y_i が標本の要素となるということを意味している。

x_i, y_i について押さえておかねばならないのは、そのサンプリングで行われているのは、母集団の、ある x_i についての y_i が選ばれるということだ。すなわち、y_i は確率変数だが、x_i は確率変数ではなく決まった値をもつということだ。決まった x_i に対して、生じうる y_i がどのようなものかを表すのがこの式なのだ。

式の中の b_0, b_1, e_i は確率変数であり、どのような標本になるかでそれらの値は変わってくる。b_1 を標本回帰係数（傾き）という。また、b_0 をここでは標本回帰の切片と呼ぶことにする。b_0, b_1 を切片、傾きと呼ぶのは、これらの値が後で述べる標本回帰直線の式の切片と傾きになるからである。y_i は b_0, b_1, e_i という確率変数と決まった値 x_i から構成される確率変数である。

2）標本における y_i の予測値の式

標本における y_i を予測する式は次のものだ（ここでの「 ^ 」は、予測値という意味）。

$$\hat{y}_i = b_0 + b_1 x_i$$

b_0, b_1 は得られる標本をもとに最小二乗法で導き出される。b_0, b_1 は得られる標本によって値が変わる確率変数であり、$\hat{y}_i$ も確率変数である。

3）残差

標本における y_i とその予測値である $\hat{y}_i$ の差を残差という。残差は次の式で表される。残差もまた得られる標本によって値が変わる確率変数である。

$$e_i = y_i - \hat{y}_i$$

また、最小二乗法が採用されているところから次のようになる（解説は次節）。

$$\sum e_i = 0$$
$$\sum x_i e_i = 0$$

4）残差の分散

回帰分析では残差の分散が重要になる。残差の分散は残差の平方和を n−2 で割ったものだ。通常の分散は n で割り、いわゆる不偏分散は n−1 で

割る。この残差の分散は n−2 で割るのである。

$$V_e = \frac{S_e}{n-2}$$
$$= \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y})^2}{n-2} = \frac{\sum (y_i - b_0 - b_1 x_i)^2}{n-2}$$

どうして n−2 で割るのかには訳があるのだが、ここでは触れない。解説は次の節に譲る。

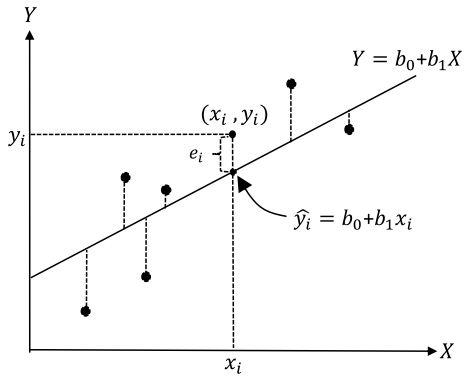
5）XY 軸グラフの標本回帰直線の式

さて、母回帰式と同様、標本回帰式についてもグラフで説明するとわかりやすいことが多い。ここでは大文字の X と Y を用いた次の標本回帰直線を使うことにする。

$$Y = b_0 + b_1 X$$

この式は $\hat{y}_i = b_0 + b_1 x_i$ をもとにした式である。もとの式は、x_i に対応する予測値の点を表すものだが、この標本回帰直線の式はその点の集合を直線で表すものだ（図 20）。

図 20　標本回帰直線と yi の分布



（4）母回帰式と標本回帰式に関わる注意点

1）回帰式と関連するいくつかの式

回帰分析では回帰式に関連したいくつかの式が現れ、混乱が生じるおそれがあるので表 13 にまとめておく。テキストによっては期待値や予測値の式､回帰直線の式を回帰式という場合もあるが、本稿でいう回帰式とは表にある回帰式である。期待値や予測値の式について、本稿では回帰の期待値の式、回帰の予測値の式などと表現する。

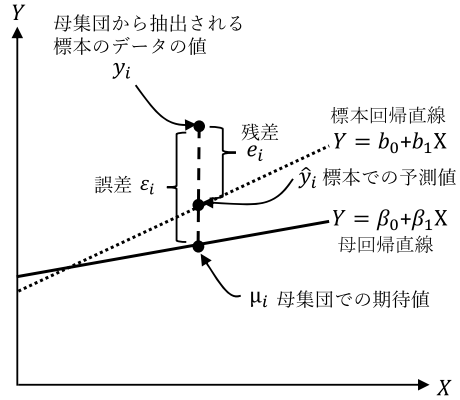
表 13　回帰分析における回帰式と関連する諸式

母集団	回帰式	$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
	期待値μ_iの式	$\mu_i = \beta_0 + \beta_1 x_i$
	回帰直線の式	$Y = \beta_0 + \beta_1 X \quad (\mu_i \in Y, x_i \in X)$
標本	回帰式	$y_i = b_0 + b_1 x_i + e_i$
	予測値y_iの式	$\hat{y}_i = b_0 + b_1 x_i$
	回帰直線の式	$Y = b_0 + b_1 X \quad (\hat{y}_i \in Y, x_i \in X)$

2）残差と誤差

母集団における誤差と標本における残差については混乱することがある。図 21 をもとに、この違いをしっかり理解しておいてほしい。

図 21　残差と誤差



ここでの母回帰直線は$\mu_i=\beta_0+\beta_1 x_i$をもとにしたものであり、標本回帰直線は$\hat{y}_i=b_0+b_1 x_i$をもとにしたものである。母集団の分布が定まっている限り、母回帰直線は決まったものになる。ただしそれは通常不可知な直線である。標本回帰直線は標本の採り方によって切片や傾きが確率的に変わる直線である。

誤差ε_iとは抽出されるy_iとその母集団での期待値μ_iとのズレであり、残差e_iとは抽出されるy_iと標本での予測値$\hat{y}_i$とのズレである。

この図では、母回帰直線とそこにあるμ_iだけが決まった線や値であり、標本回帰直線や$y_i, \hat{y}_i, \varepsilon_i, e_i$の値は確率的に変化する。

3）母回帰式と標本回帰式の確率変数

これまで、母回帰式ならびに標本回帰式にある確率変数と決まった値についてくどくど述べてきたが、まとめると表 14 のようになる。これらに付け加えて、母集団の誤差（確率変数）ε_iについて、$\varepsilon_i \sim N\left(0,\sigma^2\right)$であることを押さえておかねばならない。この$\sigma^2$は母数であり、もちろん決まった値である。

表 14　回帰分析における確率変数

	母回帰式	標本回帰式
式	$y_i=\beta_0+\beta_1 x_i+\varepsilon_i$ $\left(\mu_i=\beta_0+\beta_1 x_i\right)$	$y_i=b_0+b_1 x_i+e_i$ $\left(\hat{y}_i=b_0+b_1 x_i\right)$
確率変数	y_i　ε_i	y_i　$\hat{y}_i$　b_0　b_1　e_i
決まった値	μ_i　β_0　β_1　x_i	x_i

平均、平方和、偏差積和についても、それらが確率変数かどうかということを示しておこう（表15）。それぞれが確率変数になるかどうかは、それぞれの式の内部に確率変数が含まれるかどうかで決まる。

表 15　平均・平方和・偏差積和と確率変数

	平均	平方和・偏差積和
確率変数	$\bar{y}=\frac{1}{n}\sum y_i$	$S_{yy}=\sum\left(y_i-\bar{y}\right)^2$ $S_{xy}=\sum\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)$ $S_R=\sum\left(\hat{y}_i-\bar{y}\right)^2$ $S_e=\sum\left(y_i-\hat{y}_i\right)^2$
決まった値	$\bar{x}=\frac{1}{n}\sum x_i$	$S_{xx}=\sum\left(x_i-\bar{x}\right)^2$

4）yi の期待値、y の平均、y の平均の期待値

y_iの期待値、y の平均、y の平均の期待値はそれぞれ異なったものだ。y_iの期待値を母集団と標本の式で表現すると次のようになる。

$$E\left(y_i\right)=E\left(\beta_0+\beta_1 x_i+\varepsilon_i\right)=\beta_0+\beta_1 x_i$$
$$E\left(y_i\right)=E\left(b_0+b_1 x_i+e_i\right)=E\left(b_0\right)+E\left(b_1\right)x_i$$

これらは期待値なので決まった値になる。

y の平均を母集団と標本の式で表現すると次のようになる。

$$\bar{y}=\frac{1}{n}\sum y_i=\frac{1}{n}\sum\left(\beta_0+\beta_1 x_i+\varepsilon_i\right)$$
$$=\beta_0+\beta_1\bar{x}+\bar{\varepsilon}$$
$$\bar{y}=\frac{1}{n}\sum y_i=\frac{1}{n}\sum\left(b_0+b_1 x_i+e_i\right)$$
$$=b_0+b_1\bar{x}$$

y の平均$\bar{y}$は確率変数であり、母集団の式ではε_iによって、標本の式ではb_0とb_1によって確率的に値が変わる。$\bar{y}$は、母集団のy_iの予測値$\hat{y}_i$を用いた$\left(\sum\hat{y}_i\right)/n$でもないし、$y_i$の期待値$\mu_i$を用いた$\left(\sum\mu_i\right)/n$でもない。

y の平均の期待値を母集団と標本の式で表現すると次のようになる。期待値なのでこれらは決ま

った値になる。

$$E(\bar{y}) = E\left(\frac{1}{n}\sum y_i\right) = E(\beta_0 + \beta_1\bar{x} + \bar{\varepsilon})$$
$$= \beta_0 + \beta_1\bar{x}$$
$$E(\bar{y}) = E\left(\frac{1}{n}\sum y_i\right) = E(b_0 + b_1\bar{x})$$
$$= E(b_0) + E(b_1)\bar{x}$$

5.3 標本統計量と母数の関係

（1）母数の推定量

1）母回帰係数の推定量

母回帰係数は不可知である。そこで、標本回帰係数で母回帰係数を推定する。標本回帰係数は母回帰係数の不偏推定量である。

$$E(b_1) = \beta_1$$

2）母回帰の切片の推定量

母回帰の切片もまた不可知である。そこで、標本回帰の切片で母回帰の切片を推定する。標本回帰の切片は母回帰の切片の不偏推定量である。

$$E(b_0) = \beta_0$$

3）誤差分散σ^2の推定量

母集団の誤差ε_iの分散σ^2は通常知ることのできない値である。そこで、標本統計量である残差の分散V_eを用いてσ^2を推定する。残差の分散V_eは次の式で得られる。

$$V_e = \frac{S_e}{n-2} = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y})^2}{n-2}$$

残差の分散V_eは母集団の誤差ε_iの分散σ^2の不偏推定量である。

$$E(V_e) = \sigma^2$$

（2）標本統計量の分散と標準誤差

1）標本回帰係数の分散と標準誤差

標本回帰係数の分散はxの平方和と母集団の誤差をもとに、次のように表せることがわかっている。

$$V(b_1) = \frac{\sigma^2}{S_{xx}}$$

この分散の平方根である標準偏差は、母数を推定する標本統計量の標準偏差であるから、標本回帰係数の標準誤差と呼ばれる。標本回帰係数の標準誤差は次のものである。

$$\sqrt{V(b_1)} = \sqrt{\frac{\sigma^2}{S_{xx}}}$$

ところで、式にあるσ^2は不可知なので、そのかわりにσ^2の不偏推定量V_eを用いることがある。これも標本回帰係数の標準誤差である。具体的な標本にもとづく標準誤差の計算はこちらが用いられる。

$$\sqrt{V(b_1)} = \sqrt{\frac{V_e}{S_{xx}}}$$

2）標本回帰の切片の分散と標準誤差

標本回帰の切片の分散は次のようになる。

$$V\left(b_0\right)=\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{\mathrm{S}_{xx}}\right)$$

この分散の平方根である標準偏差は、母数を推定する標本統計量の標準偏差であるから、切片の標準誤差と呼ばれる。切片の標準誤差は次のものである。

$$\sqrt{V\left(b_0\right)}=\sqrt{\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{\mathrm{S}_{xx}}\right)}$$

式にあるσ^2は不可知なので、そのかわりにσ^2の不偏推定量V_eを用いることがある。これも切片の標準誤差である。具体的な標本にもとづく標準誤差の計算はこちらが用いられる。

$$\sqrt{V\left(b_0\right)}=\sqrt{V_e\left(\frac{1}{n}+\frac{\bar{x}^2}{\mathrm{S}_{xx}}\right)}$$

表 16 は以上のことをまとめたものである。

表 16　回帰分析における母数の推定量

母数	推定量	期待値	標準誤差
β_1	$b_1=\frac{S_{xy}}{S_{xx}}$	$E\left(b_1\right)=\beta_1$	$\sqrt{V\left(b_1\right)}=\sqrt{\frac{\sigma^2}{\mathrm{S}_{xx}}}$ $\sqrt{V\left(b_1\right)}=\sqrt{\frac{V_e}{\mathrm{S}_{xx}}}$
β_0	$b_0=\bar{y}-\frac{S_{xy}}{S_{xx}}\bar{x}$	$E\left(b_0\right)=\beta_0$	$\sqrt{V\left(b_0\right)}=\sqrt{\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{\mathrm{S}_{xx}}\right)}$ $\sqrt{V\left(b_0\right)}=\sqrt{V_e\left(\frac{1}{n}+\frac{\bar{x}^2}{\mathrm{S}_{xx}}\right)}$
σ^2	$V_e=\frac{S_e}{n-2}$	$E\left(V_e\right)=\sigma^2$	

5.4 標本統計量の確率分布と検定・推定

（1）回帰係数の確率分布と検定・推定

1）確率分布と検定

回帰分析の仮定が満たされているとき、標本回帰係数は上でみた期待値（β_1）と分散（σ^2/S_{xx}）を持つ正規分布に従う。

$$b_1\sim N\left(\beta_1,\frac{\sigma^2}{S_{xx}}\right)$$

この式のb_1から分布の平均β_1を引き、標準偏差$\sqrt{\sigma^2/S_{xx}}$で割って標準化すると、次のように標準正規分布に従う式が導ける。

$$\frac{b_1-\beta_1}{\sqrt{\frac{\sigma^2}{\mathrm{S}_{xx}}}}\sim N\left(0,1\right)$$

さて、ここで式の中にあるσ^2は通常不可知である。そこでσ^2のかわりに、その推定値であるV_eを式に入れる。推定値であるV_eを式に入れることによって、この式はもはや標準正規分布には従わなくなる。それは自由度 n−2 の t 分布に従うのである。

$$\boxed{\frac{b_1-\beta_1}{\sqrt{\frac{V_e}{\mathrm{S}_{xx}}}}\sim t_{n-2}}$$

これは母回帰係数がある値 a であるという帰無仮説$\left(H_0:\beta_1=a\right)$の検定に用いる検定統計量である。

母回帰係数が 0 であるという帰無仮説$\left(H_0:\beta_1=0\right)$の検定には、この式の$\beta_1$に 0 を代入した次の検定統計量を用いる。

$$Test\left(H_0:\beta_1=0\right)=\frac{b_1}{\sqrt{\frac{V_e}{\mathrm{S}_{xx}}}}\sim t_{n-2}$$

2）95%の信頼区間

自由度 n－2 の t 分布の上側確率 2.5%に対応する値を$t_{n-2}\left(0.025\right)$で表すとき、回帰係数の 95%の信頼区間は次のようになる。

$$b_1-t_{n-2}\left(0.025\right)\sqrt{\frac{V_e}{\mathrm{S}_{xx}}}\le\beta_1\le b_1+t_{n-2}\left(0.025\right)\sqrt{\frac{V_e}{\mathrm{S}_{xx}}}$$

99%の信頼区間を求めたければ、次の式を使う。

$$b_1-t_{n-2}\left(0.005\right)\sqrt{\frac{V_e}{\mathrm{S}_{xx}}}\le\beta_1\le b_1+t_{n-2}\left(0.005\right)\sqrt{\frac{V_e}{\mathrm{S}_{xx}}}$$

ここで 5.1 （4）の「検定と区間推定」の話を思い出してほしいのだが、現実の標本の値を用いた信頼区間の式は、当たっているかはずれているかのどちらかであり、確率の式とはいえない。信頼区間の「信頼度 95%」とは、さまざまな現実の標本で信頼区間を算出した時、それらのうちの 95%は正しい式になるという意味である。

（2）回帰係数の検定と相関係数の検定

ところで、回帰係数と相関係数の間には次のような関係がある。

$$b_1=\left(\frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}}\right)r\qquad r=\left(\frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}}\right)b_1$$

ここから、x の分散と y の分散が等しいときには、相関係数の値と回帰係数の値は等しくなることがわかる。したがって、変数 x と変数 y がともに標準化されている際には、相関係数は回帰係数と等しい。

また、相関係数の検定（$H_0:\rho=0$）は、回帰係数の検定（$H_0:\beta_1=0$）と同じになることも知っておいてほしい。相関の検定では次の検定統計量が用いられる。

$$Test\left(H_0:\rho=0\right)=\frac{r}{\sqrt{\left(1-r^2\right)/\left(n-2\right)}}\sim t_{n-2}$$

他方、上で見たように、回帰係数の検定は次の検定統計量を用いる。

$$Test\left(H_0:\beta_1=0\right)=\frac{b_1}{\sqrt{V_e/S_{xx}}}\sim t_{n-2}$$

証明は後の節に譲るが、両者は変形していくと同じ式になるのである。

$$\frac{r}{\sqrt{\left(1-r^2\right)/\left(n-2\right)}}=\frac{b_1}{\sqrt{V_e/S_{xx}}}$$

したがって、2 つの帰無仮説の検定は同等である。

（3）切片の確率分布と検定・推定

1）確率分布と検定

回帰分析の仮定が満たされているとき、標本回帰の切片は上でみた期待値と分散を持つ正規分布に従うことになる。

$$b_1 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)\right)$$

この式のb_0から分布の平均β_0を引き、標準偏差$\sqrt{\sigma^2\left(1/n+\bar{x}^2/S_{xx}\right)}$で割って標準化すると、次のように標準正規分布に従う式が導ける。

$$\frac{b_0-\beta_0}{\sqrt{\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}} \sim N(0,1)$$

ここで式の中にあるσ^2は不可知なので、その推定値であるV_eを式に入れる。推定値であるV_eを式に入れることによって、この式はもはや標準正規分布には従わなくなる。それは自由度 n－2 の t 分布に従うことになる。

$$\boxed{\frac{b_0-\beta_0}{\sqrt{V_e\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{n-2}}$$

これは母回帰の切片がある値 a であるという帰無仮説（$H_0:\beta_0=a$）の検定で用いる検定統計量である。

$\beta_0=0$の検定には次の検定統計量を用いる。

$$\frac{b_0}{\sqrt{V_e\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{n-2}$$

2）95%の信頼区間

自由度 n－2 の t 分布の上側確率 2.5%に対応する値を$t_{n-2}(0.025)$で表すとき、切片の 95%の信頼区間は次のようになる。

$$b_0 - t_{n-2}(0.025)\sqrt{V_e\left(\frac{1}{n}+\frac{\bar{x}^2}{S_x}\right)} \le \beta_0 \le b_0 + t_{n-2}(0.025)\sqrt{V_e\left(\frac{1}{n}+\frac{\bar{x}^2}{S_x}\right)}$$

99%の信頼区間は次のとおりだ。

$$b_0 - t_{n-2}(0.005)\sqrt{V_e\left(\frac{1}{n}+\frac{\bar{x}^2}{S_x}\right)} \le \beta_0 \le b_0 + t_{n-2}(0.005)\sqrt{V_e\left(\frac{1}{n}+\frac{\bar{x}^2}{S_x}\right)}$$

以上で、回帰係数と切片の確率分布について理解できたはずだ。まとめると表 17 のようになる。

表 17　回帰係数と切片の確率分布

	確率分布	標準正規分布との関係 t分布との関係
回帰係数 b_1	$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$	$\frac{b_1-\beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1)$ $\frac{b_1-\beta_1}{\sqrt{\frac{V_e}{S_{xx}}}} \sim t_{n-2}$
切片 b_0	$b_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)\right)$	$\frac{b_0-\beta_0}{\sqrt{\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}} \sim N(0,1)$ $\frac{b_0-\beta_0}{\sqrt{V_e\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{n-2}$

(4) 回帰モデルの検討

1）全平方和・残差平方和・回帰平方和

回帰分析においては平方和や分散をもとに、推

定された回帰の予測式に意味があるか、それは何らかの説明をしているのかどうかを検討することがある。これを回帰モデルの検討という。その際に考察のもととなる式は全平方和、回帰平方和、残差平方和に関する次の式である。

$$S_{yy} = S_e + S_R$$

その中身は次のようになっている。

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

すなわち、「全平方和＝残差平方和＋回帰平方和」が回帰分析の仮定のもとで成立しているのである。

2）決定係数による回帰モデルの検討

回帰式の検討には決定係数 R^2 が用いられることがある。決定係数は、上の回帰平方和を全平方和で割ったものだ。それは、標本の全変動のどの程度が回帰変動で説明されるかを意味しており、寄与率と呼ばれることもある。単回帰分析の場合、決定係数の値は相関係数の 2 乗の値と同じになる。

$$R^2 = \frac{S_R}{S_{yy}} = 1 - \frac{S_e}{S_{yy}}$$

回帰モデルが母集団の状況を説明しているのかどうかの検討には下の統計量が用いられる。

$$\frac{R^2}{\left(1-R^2\right)/\left(n-2\right)} \sim F^1_{n-2}$$

この統計量は、当該回帰モデルが母集団で何の説明力も持たない場合、自由度（1,n−2）の F 分布に従う。

ところで、相関係数の検定統計量（$H_0 : \rho = 0$）は下のものであり、それは自由度 n−2 の t 分布に従う。

$$Test\left(H_0 : \rho = 0\right) = \frac{r}{\sqrt{\left(1-r^2\right)/\left(n-2\right)}} \sim t_{n-2}$$

単回帰モデルでは、決定係数は相関係数の 2 乗となるが、この統計量の 2 乗が 2 つ上の式で示された回帰モデル検討のための統計量になる（t 分布に従う変数を 2 乗するとその変数は F 分布に従うので、この統計量は F 分布に従うことになる）。

3）分散比による回帰モデルの検討

回帰に意味があるのかどうかということは、次の式で表される分散比で検討されることもある。

$$\frac{S_R}{V_e} = \frac{S_R/1}{S_e/\left(n-2\right)}$$

この式の右の項の分子は回帰平方和 S_R をその自由度 1 で割ったものであり、分母は残差平方和 S_e をその自由度 n−2 で割ったものである。これらはそれぞれ回帰の平均平方、残差の平均平方と呼ばれる。残差の平均平方は残差分散に等しい。

回帰モデルが母集団の状況をまったく説明していない場合 $\beta_1 = 0$ となるが、そのときこの分散比は自由度（1,n−2）の F 分布に従う。

$$\frac{S_R/1}{S_e/\left(n-2\right)} \sim F^1_{n-2}$$

4）回帰モデルの検討と回帰係数の検定

分散比の検討は回帰係数の検定（帰無仮説は $\beta_1 = 0$）と同等である。そして、証明は次節に譲る

が、決定係数の検討で用いる式と分散比の検討で用いる式は変形していくと同じになる（下の等式が成り立つ）。

$$\frac{R^2}{\left(1-R^2\right)/\left(n-2\right)}=\frac{S_R/1}{S_e/\left(n-2\right)}$$

また、母回帰係数の検定（$\beta_1=0$）は相関係数の検定（$\rho=0$）と同等であることを前に述べた。以上のことを全部まとめると、結局、決定係数の検討、分散比の検討、回帰係数の検定、相関係数の検定はすべて同じことをしていることになる。

ただし、このことは説明変数が 1 つの単回帰分析においてのみいえることで、説明変数が 2 つ以上の重回帰分析の場合は様子が少し変わってくる（そこでは回帰係数は退出し偏回帰係数が登場する。また、重相関係数というものが重要な役割を果たすようになる。そして決定係数の検討と分散比の検討と重相関係数の検定は同じものとなり、偏回帰係数の検定はそれらとは異なったものになる）。

5）回帰分析の前提となる仮定の検討

推測統計的な回帰分析においては、母集団における対象の分布が図 16 のイメージのようになっているという仮定が設けられている。すなわち、X と Y の間には直線的な関係があり、Y の分散はどの X の点においても等しくなければ、推測統計的な回帰分析の議論は成り立たない。

これらは母集団の話であるので直接検討することはできない。検討に際しては、標本の散布図などを見て、仮定から大きくずれていないかを間接的に探るという方法などがとられる。

5.5 回帰分析の実例 2

（1）データと基本統計量

以上の知識をもとに回帰分析の実際の分析結果を見てみよう。ここで用いるのは、3 節で見た表 2 の都道府県データである。本来、この都道府県データは全数調査のデータと見なされるので検定は不要と考えられるのだが、ここではそのデータを何らかの母集団から得られたものと想定し、分析結果を検討する。

データの基本統計量と、変数の相互関係を再度示しておこう（表 18）。

表 18　県別データの基本統計量・相互関係

	自殺率	高齢化率	失業率	県民所得
平均	18.845	28.285	3.051	2.874
偏差平方和	237.836	352.220	16.837	11.810
分散	5.060	7.494	0.358	0.251
標準偏差	2.250	2.738	0.599	0.501

		自殺率	高齢化率	失業率	県民所得
自殺率	偏差積和S_{xy}	237.836	92.811	8.323	-11.461
	共分散s_{xy}	5.060	1.975	0.177	-0.244
	相関係数r_{xy}	1	0.321	0.132	-0.216
	有意確率		0.028	0.378	0.144
高齢化率	偏差積和S_{xy}	92.811	352.220	77.010	-26.445
	共分散s_{xy}	1.975	7.494	1.639	-0.563
	相関係数r_{xy}	0.321	1	-0.362	-0.410
	有意確率	0.028		0.012	0.004
失業率	偏差積和S_{xy}	8.323	77.010	16.837	-2.507
	共分散s_{xy}	0.177	1.639	0.358	-0.053
	相関係数r_{xy}	0.132	-0.362	1	-0.178
	有意確率	0.378	0.012		0.232
県民所得	偏差積和S_{xy}	-11.461	-26.445	-2.507	11.810
	共分散s_{xy}	-0.244	-0.563	-0.053	0.251
	相関係数r_{xy}	-0.216	-0.410	-0.178	1
	有意確率	0.144	0.004	0.232	

（2）回帰係数の表

完全失業率 X で自殺率 Y を説明する回帰分析を例にして解説しよう。表 19 は、モデルの回帰

係数・切片の結果を示したものである。統計ソフトによって多少の違いはあるが、回帰分析をおこなった場合は、まずこのような情報が結果として示される。

表 19　回帰係数の表（失業率→自殺率）

	非標準化係数	標準誤差	標準化係数	t	有意確率
失業率	0.494	0.555	0.132	0.890	0.378
［定数］	17.337	1.727		10.039	0.000
決定係数	0.017				

表から、標本回帰直線が次の式で表せることがわかる。

$$Y = 17.337 + 0.494X$$

また、データが標準化されている場合は次のようになる。

$$Y = 0.132X$$

表には、決定係数は 0.017 であることも示されている。以上は 3.4 で説明した通りだ。

次の列の「標準誤差」に示されているのは、回帰係数と切片の標準誤差の値である、これらは次の式で得られたものだ。

$$\sqrt{V(b_1)} = \sqrt{\frac{V_e}{S_{xx}}}$$

$$\sqrt{V(b_0)} = \sqrt{V_e\left(\frac{1}{n} + \frac{\bar{x}^2}{\mathrm{S}_{xx}}\right)}$$

式の中にある S_{xx}、$\bar{x}$ の値は表 18 にあり、それぞれ 16.837、3.051 である。V_e の値は、やや先走るが表 20 の残差の平均平方（残差分散）のことであり、5.194 となっている。これらを用いると、

$$\sqrt{V(b_1)} = \sqrt{\frac{V_e}{S_{xx}}} = \sqrt{\frac{5.194}{16.837}} = 0.555$$

$$\sqrt{V(b_0)} = \sqrt{V_e\left(\frac{1}{n} + \frac{\bar{x}^2}{\mathrm{S}_{xx}}\right)}$$

$$= \sqrt{5.194\left(\frac{1}{47} + \frac{3.051^2}{16.837}\right)} = 1.727$$

となって回帰係数の表（表 19）の値と一致する。

次の列の「t」は、母回帰直線の回帰係数 $\beta_1 = 0$、切片 $\beta_0 = 0$ という帰無仮説が正しい場合に、検定統計量がどうなるかを示したものである。「失業率」の行に示されているのが回帰係数についての検定統計量であり、これは標準誤差を用いて、

$$Test = \frac{b_1}{\sqrt{V(b_1)}}$$

$$= \frac{b_1}{\sqrt{V_e / S_{xx}}} = \frac{0.494}{0.555} = 0.890$$

のようにして求められたものだ。

同様に「定数」の行に示されているのが切片についての検定統計量であり、

$$Test = \frac{b_0}{\sqrt{V(b_0)}}$$

$$= \frac{b_0}{\sqrt{V_e\left(\frac{1}{n} + \frac{\bar{x}^2}{\mathrm{S}_{xx}}\right)}} = \frac{17.337}{1.727} = 10.039$$

のように求められている。

「有意確率」は、上記の帰無仮説が正しい場合

に、t 分布において左の検定統計量の値が出る確率を意味する（きちんと言えば、その値に正負の記号をつけたとき、正負のその値より数直線上の外側の値が出る確率の計）。この値が 5%以下だと、通常、母回帰直線における回帰係数 β_1 は 0 ではない、切片 β_0 は 0 ではないと判断される。

今回の失業率から自殺率を予測する回帰分析では、有意確率は 0.378 で、有意水準 5%（0.05）を超えている。したがって、母回帰係数 $\beta_1 = 0$ という帰無仮説は棄却できないという結果になる。ここから「母集団において、失業率が自殺率に影響力を持つとは言えない」という結論が導ける。「持つとは言えない」であって「持たない」ではないということに注意してほしい。

今回の分析において、切片についての帰無仮説、$\beta_0 = 0$ は棄却できる。ただこれが棄却できても失業率と自殺率の関係について、あまり付加的な情報は得られない。切片の検定は、通常ほとんど利用されることはない。

（3）平方和の表

統計ソフトでは、次のような表も出力されるはずだ（表 20）。これは回帰分析における平方和などの結果を示した表である。

表 20　平方和の表（失業率→自殺率）

	平方和	自由度	平均平方	F	有意確率
回帰	4.114	1	4.114	0.792	0.378
残差	233.722	45	5.194		
合計	237.836	46			

表から回帰平方和、残差平方和、全平方和がそれぞれ、

$$S_R = \sum \left(\hat{y}_i - \bar{y} \right)^2 = 4.114$$
$$S_e = \sum \left(y_i - \hat{y}_i \right)^2 = 233.722$$
$$S_{yy} = \sum \left(y_i - \bar{y} \right)^2 = 237.836$$

であることがわかる。これらの平方和をもとに決定係数は、

$$R^2 = \frac{S_R}{S_{yy}} = \frac{4.114}{237.836} = 0.0173$$

と導ける。この値は回帰係数の表（表 19）に掲載されている決定係数の値と一致する。

平方和をその右の列にある自由度で割ると、平均平方になる。残差の平均平方は残差分散に一致する。

$$V_e = \frac{S_e}{n-2} = \frac{233.722}{45} = 5.194$$

回帰の平均平方を残差の平均平方（残差分散）で割ると、表の「F」の値が出る。これは分散比とよばれるものだ。

$$\frac{S_R/1}{S_e/(n-2)} = \frac{4.114/1}{233.722/45} = \frac{4.114}{5.194} = 0.792$$

この分散比は、母回帰係数が 0 のとき、自由度（1,n-2）の F 分布に従う。さらにこの分散比は、決定係数 R^2 をもとにした下の検定統計量とも一致する。

$$\frac{R^2}{\left(1-R^2\right)/(n-2)}$$

$$= \frac{0.0173}{(1-0.0173)/(47-2)} = 0.792$$

説明変数が 1 つの単回帰分析では、回帰係数の検定、相関係数の検定、分散比を用いた検定、決定係数を用いた検定がすべて同等のものとなる。読者はそれらの有意確率がすべて同じ（0.378）になっていることを確認するといい。

（4）信頼区間

標準誤差の情報を用いれば、回帰係数や切片の信頼区間を求めることができる。表 19 の回帰係数と標準誤差の値を用いて、回帰係数と切片の 95%信頼区間を求めてみよう。

回帰係数の 95%の信頼区間の式は次の通りだった。

$$b_1 - t_{n-2}(0.025)\sqrt{\frac{V_e}{S_{xx}}} \le \beta_1 \le b_1 + t_{n-2}(0.025)\sqrt{\frac{V_e}{S_{xx}}}$$

したがって、信頼区間は次のようになる。

$$0.494 - 2.014 \times 0.555 \le \beta_1 \le 0.494 + 2.014 \times 0.555$$
$$\therefore -0.624 \le \beta_1 \le 1.612$$

切片の 95%信頼区間の式は次のとおりである。

$$b_0 - t_{n-2}(0.025)\sqrt{V_e\left(\frac{1}{n} + \frac{\bar{x}^2}{S_x}\right)} \le \beta_0 \le b_0 + t_{n-2}(0.025)\sqrt{V_e\left(\frac{1}{n} + \frac{\bar{x}^2}{S_x}\right)}$$

したがって、信頼区間は次のようになる。

$$17.337 - 2.014 \times 1.727 \le \beta_0 \le 17.337 + 2.014 \times 1.727$$
$$\therefore 13.859 \le \beta_0 \le 20.815$$

なお 5.1 （4）で述べた信頼区間と検定結果の関係は、上記のようにして求められる回帰係数や切片の場合でも同じだ。すなわち、表 19 では、母回帰係数 $\beta_1 = 0$ という帰無仮説が有意水準 5%（両側）で棄却できないことが示されているが、そのことはここで求めた「母回帰係数の 95%の信頼区間の範囲に 0 が含まれている」ということに対応しているのである。

5.6 課題

（1）問題

都道府県データを用い、高齢化率 X から県民所得 Y を予測するような回帰分析を行ったところ、表 21 および表 22 の結果が出力された。

表 21　回帰係数の表（高齢化率→県民所得）

	非標準化係数	標準誤差	標準化係数	t	有意確率
高齢化率	-0.075	0.025	-0.410	-3.016	0.004
［定数］	4.998	0.708		7.064	0.000
決定係数	0.168				

表 22　平方和の表（高齢化率→県民所得）

	平方和	自由度	平均平方	F	有意確率
回帰	1.986	1	1.986	9.095	0.004
残差	9.825	45	0.218		
合計	11.810	46			

これらの表を見て以下の問いに答えよ。なお母回帰直線の切片を β_0、回帰係数を β_1、標本回帰

直線の切片を b_0、回帰係数 b_1 をとする。また、検定の有意水準は5%に設定すること。

a) 回帰直線の式、ならびに標準得点にもとづく回帰直線の式を書け。また、これらは母回帰直線の式か、標本回帰直線の式か述べよ。
b) 回帰係数の検定に関する帰無仮説、切片の検定に関する帰無仮説は何か。
c) 回帰係数の行にある有意確率は、両側検定の結果である。この表で示されていることは、t 分布で−3.016 よりも小さな値が出る確率は 0.004 ということか。
d) この t 分布の自由度を述べよ。
e) 回帰係数に関する帰無仮説、切片に関する帰無仮説はそれぞれ棄却できるか判断せよ。
f) 回帰係数の帰無仮説の判断の結果どのような結論が導かれるか。
g) この結果から高齢の者ほど所得が下がる傾向があると言えるか。
h) 表に示されている標準誤差の値を用いて、回帰係数の 95%信頼区間、切片の 95%信頼区間を求めよ。
i) 回帰係数について、信頼区間と検定結果の対応関係について説明せよ。
j) 回帰平方和、残差平方和、全平方和の値を用いて決定係数を計算し、回帰係数の表にある数値と一致していることを確認せよ。
k) 平方和の表にある有意確率は、F 分布で 9.095 以上の値が出る確率は 0.004 ということか。
l) この F 分布の自由度を述べよ。
m) 平方和の表に検定結果が記載されているが、この検定の帰無仮説は何か。
n) 回帰係数の表の t の値−3.016 を 2 乗すると平方和の表の F の値 9.095 に非常に近い値になる。このことは何を意味しているのか。

（2）解答

a) $Y = 4.998 - 0.075X$ （非標準得点にもとづく）
$Y = -0.410X$ （標準得点にもとづく）。ともに標本回帰直線の式。
b) 回帰係数に関する帰無仮説： $\beta_1 = 0$、切片に関する帰無仮説： $\beta_0 = 0$
c) 違う。−3.016 と 3.016 の外側の値が出る確率が 0.004 ということである。
$P\left(|t| \geq 3.016\right) = 0.004$
d) $n - 2 = 47 - 2 = 45$
e) 回帰係数 : 有意水準 5%で帰無仮説は棄却できる。切片 : 有意水準 5%で帰無仮説は棄却できる。
f) 母集団においても県の高齢化率は県の 1 人あたり県民所得に対して（負の）影響力を持つ。
g) このデータは都道府県レベルのデータであるから、個人について「高齢になると収入が下がる」といったとは言えない。言えるのは「高齢県になると低県民所得県になる」といったと都道府県レベルのことだけである。
h) 回帰係数： $-0.125 \leq \beta_1 \leq -0.025$
切片： $3.573 \leq \beta_0 \leq 6.423$
i) 求められた信頼区間に 0 は含まれない。これに対応して、回帰係数の検定では、有意水準 5%で帰無仮説が棄却されている。
j)
$$R^2 = \frac{S_R}{S_{yy}} = \frac{1.986}{11.810} = 0.168$$
k) そうである。
l) $\left(1, n-2\right) = \left(1, 45\right)$
m) 分散比を用いた検定の帰無仮説も、回帰係数の検定と同じ $\beta_1 = 0$ である。
n) 自由度 n−2 の t 分布に従う確率変数の 2 乗は自由度（1,n−2）の F 分布に従うので、確率についての下の関係が成り立つ。
$$P\left(|t_{45}| \geq 3.016\right) \Leftrightarrow P\left(F_{45}^{1} \geq 3.016^2\right)$$

6　回帰分析の数学的構造

6.1 全体像の把握

回帰分析は以下の仮定をもとにした分析法であり、ここからすべてが始まる。

$$\begin{cases} y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \\ y_i = b_0 + b_1 x_i + e_i \\ \varepsilon_i \sim N\left(0, \sigma^2\right) \\ Cov\left(\varepsilon_i, \varepsilon_j\right) = 0 \quad (i \neq j) \\ \text{最小二乗法による} b_0 \text{と} b_1 \text{の推定} \end{cases}$$

これらの仮定は次のことをも意味している。

$$E\left(\varepsilon_i\right) = 0$$
$$V\left(\varepsilon_i\right) = \sigma^2$$
$$E\left(\varepsilon_i^2\right) = \sigma^2$$
$$E\left(\bar{\varepsilon}\right) = 0$$
$$\hat{y}_i = b_0 + b_1 x_i$$
$$E\left(y_i\right) = \mu_i = \beta_0 + \beta_1 x_i$$

そして、最小二乗法を採用しているところから次の式が直接導ける。

$$\sum e_i = \sum \left(b_0 + b_1 x_i - y_i\right) = 0$$
$$\sum x_i e_i = \sum x_i \left(b_0 + b_1 x_i - y_i\right) = 0$$

また、次のことも仮定や最小二乗法の採用から導ける。

$$\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{\varepsilon}$$

$$\bar{y} = b_0 + b_1 \bar{x}$$
$$\bar{e} = \frac{\sum e_i}{n} = 0$$

そして、これらをもとに前節の表 17 のようなことが導き出せ、それらにもとづいて標本から母集団の状態を推測していくのである。ではどのようにして表 17 のようなことが導けるのか。以下、順を追って解説していく[3]。

6.2 回帰係数と切片の最小二乗推定

（1）残差平方和の式と２つの導関数

実現値 y_i についての予測値 $\hat{y}_i$ の式は次のものである。

$$\hat{y}_i = b_0 + b_1 x_i$$

したがって、実現値 y_i と予測値 $\hat{y}_i$ の差（残差）e_i と、その差の 2 乗の総和である S_e（残差平方和）は下の式で表される。

$$e_i = y_i - (b_0 + b_1 x_i)$$
$$S_e = \sum \left(b_0 + b_1 x_i - y_i\right)^2$$

最小二乗法はこの残差平方和 S_e が最小になるように b_0, b_1 を設定する方法だ。

S_e が最小であるとき下の 2 つの導関数はともに 0 である。

[3] 以下の記述では、永田・棟近（2001）、小寺（1986）、鈴木・山田（2004）等の証明を参考にしている。

$$\begin{cases}\dfrac{\partial S}{\partial b_0}=\dfrac{\partial}{\partial b_0}\sum\left(b_0+b_1x_i-y_i\right)^2=0\\ \dfrac{\partial S}{\partial b_1}=\dfrac{\partial}{\partial b_1}\sum\left(b_0+b_1x_i-y_i\right)^2=0\end{cases}$$

ここで、高校の数Ⅲで学ぶ微分の次のような公式を紹介しよう。

$$\left\{g\left(f\left(x\right)\right)\right\}'=g'\left(f\left(x\right)\right)f'\left(x\right)$$

すなわち、合成関数の微分は、それを構成するそれぞれの関数の微分の積になっている。たとえば、次のようになる。

$$h(x)=\left(3x^2+5x\right)^4$$
$$f\left(x\right)=\Delta=3x^2+5x$$
$$f'\left(x\right)=6x+5$$
$$g\left(\Delta\right)=\Delta^4$$
$$g'\left(\Delta\right)=4\Delta^3$$
$$h'(x)=g'\left(f\left(x\right)\right)f'\left(x\right)=4\left(\Delta\right)^3\left(6x+5\right)$$
$$=4\left(3x^2+5x\right)^3\left(6x+5\right)$$

この知識と、「シグマの微分は微分のシグマ」という知識をもとに、さきの上の式を微分すると次のようになる。

$$\frac{\partial S}{\partial b_0}=\frac{\partial}{\partial b_0}\sum\left(b_0+b_1x_i-y_i\right)^2$$
$$=\sum 2\left(b_0+b_1x_i-y_i\right)=0$$
$$\therefore nb_0+b_1\sum x_i=\sum y_i$$

このとき、途中の式から次のようになっていることにも注目してほしい。これは最小二乗法が使われていることによる直接の帰結である。

$$\boxed{\sum e_i=\sum\left(b_0+b_1x_i-y_i\right)=0}$$

したがって、

$$\boxed{\overline{e}=\frac{\sum e_i}{n}=0}$$

次に、さきの下の式の微分をしよう。

$$\frac{\partial S}{\partial b_1}=\frac{\partial}{\partial b_1}\sum\left(b_0+b_1x_i-y_i\right)^2$$
$$=\sum 2\left(b_0+b_1x_i-y_i\right)x_i$$
$$=2\sum\left(b_0x_i+b_1x_i^2-x_iy_i\right)=0$$
$$\therefore b_0\sum x_i+b_1\sum x_i^2=\sum x_iy_i$$

このとき、途中の式から次のようになっていることにも注目してほしい。これもまた最小二乗法が採用されていることによる直接の帰結だ。

$$\boxed{\sum x_ie_i=\sum x_i\left(b_0+b_1x_i-y_i\right)=0}$$

さて、偏微分によって下の 2 つの式が得られた。この連立方程式を回帰方程式という。

$$\begin{cases}nb_0+b_1\sum x_i=\sum y_i\\ b_0\sum x_i+b_1\sum x_i^2=\sum x_iy_i\end{cases}$$

（2）回帰方程式の解法

上式より得られる

$$b_0=\frac{\sum y_i-b_1\sum x_i}{n}$$

を下式に代入する。

$$b_0\sum x_i + b_1\sum x_i^2 = \sum x_i y_i$$
$$\sum x_i\left(\frac{\sum y_i - b_1\sum x_i}{n}\right) + b_1\sum x_i^2 = \sum x_i y_i$$
$$\frac{\sum x_i\sum y_i - b_1\sum x_i\sum x_i}{n} + b_1\sum x_i^2 = \sum x_i y_i$$
$$\sum x_i\sum y_i - b_1\left(\sum x_i\right)^2 + nb_1\sum x_i^2 = n\sum x_i y_i$$
$$nb_1\sum x_i^2 - b_1\left(\sum x_i\right)^2 = n\sum x_i y_i - \sum x_i\sum y_i$$
$$b_1\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right) = n\sum x_i y_i - \sum x_i\sum y_i$$
$$b_1 = \frac{n\sum x_i y_i - \sum x_i\sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{n\sum x_i y_i - n\bar{x}n\bar{y}}{n\sum x_i^2 - n^2\bar{x}^2}$$
$$= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{\sum\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sum\left(x_i-\bar{x}\right)^2}$$
$$= \frac{S_{xy}}{S_{xx}} = \frac{s_{xy}}{s_x^2}$$

したがって、

$$\boxed{b_1 = \frac{S_{xy}}{S_{xx}}}$$

これを先のb_0の式に入れると

$$b_0 = \frac{\sum y_i - b_1\sum x_i}{n} = \frac{\sum y_i - \frac{S_{xy}}{S_{xx}}\sum x_i}{n}$$
$$= \frac{n\bar{y} - \frac{S_{xy}}{S_{xx}}n\bar{x}}{n} = \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x} = \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x}$$

したがって、

$$\boxed{b_0 = \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x}}$$

結局、求めるy_iの予測式は次のようになる。

$$\hat{y}_i = \left(\bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x}\right) + \left(\frac{S_{xy}}{S_{xx}}\right)x_1$$

あるいは、

$$\hat{y}_i = \left(\bar{y} - \frac{s_{xy}}{s_x^2}\bar{x}\right) + \left(\frac{s_{xy}}{s_x^2}\right)x_i$$

ここで、$x_i = \bar{x}$ とすると、

$$\hat{y}_i = \left(\bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x}\right) + \left(\frac{S_{xy}}{S_{xx}}\right)\bar{x}$$
$$= \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x} + \frac{S_{xy}}{S_{xx}}\bar{x} = \bar{y}$$

となるので、次の式が成立することがわかる。

$$\boxed{\bar{y} = b_0 + b_1\bar{x}}$$

ここでついでにyの平均$\bar{y}$について、母集団からも見ておこう。

$$\bar{y} = \frac{1}{n}\sum y_i = \frac{1}{n}\sum\left(\beta_0 + \beta_1 x_i + \varepsilon_i\right)$$
$$= \frac{1}{n}\left(n\beta_0 + \beta_1\sum x_i + \sum\varepsilon_i\right)$$
$$= \frac{1}{n}\left(n\beta_0 + n\beta_1\bar{x} + n\bar{\varepsilon}\right)$$
$$= \beta_0 + \beta_1\bar{x} + \bar{\varepsilon}$$

だから

$$\boxed{\bar{y} = \beta_0 + \beta_1\bar{x} + \bar{\varepsilon}}$$

6.3 母数の推定量としての標本統計量

(1) 標本回帰係数の期待値と分散

1) 回帰係数 b_1 の期待値

次に確率変数である回帰係数がどのような期待

値と分散を持つかを明らかにしよう。まずb_1について。

$$E(b_1)=E\left(\frac{S_{xy}}{S_{xx}}\right)=\frac{E(S_{xy})}{S_{xx}}$$
$$=\frac{1}{S_{xx}}E\left(\sum(x_i-\bar{x})(y_i-\bar{y})\right)$$
$$=\frac{1}{S_{xx}}\sum(x_i-\bar{x})E(y_i-\bar{y})$$
$$=\frac{1}{S_{xx}}\sum(x_i-\bar{x})\left(E(y_i)-E(\bar{y})\right)$$
$$=\frac{1}{S_{xx}}\sum(x_i-\bar{x})\begin{pmatrix}E(\beta_0+\beta_1x_i)-\\E(\beta_0+\beta_1\bar{x})\end{pmatrix}$$
$$=\frac{1}{S_{xx}}\sum(x_i-\bar{x})\begin{pmatrix}E(\beta_0)+E(\beta_1x_i)\\-E(\beta_0)-E(\beta_1\bar{x})\end{pmatrix}$$

（β_0,β_1 や$x_i,\bar{x}$は確率変数ではないので）

$$=\frac{1}{S_{xx}}\sum(x_i-\bar{x})(\beta_0+\beta_1x_i-\beta_0-\beta_1\bar{x})$$
$$=\frac{1}{S_{xx}}\sum(x_i-\bar{x})\left(\beta_1(x_i-\bar{x})\right)$$
$$=\frac{1}{S_{xx}}\sum(x_i-\bar{x})^2\beta_1=\frac{1}{S_{xx}}S_{xx}\beta_1=\beta_1$$

したがって、

$$\boxed{E(b_1)=\beta_1}$$

2）回帰係数 b_1 の分散

$$V(b_1)=V\left(\frac{S_{xy}}{S_{xx}}\right)=\frac{1}{S_{xx}^{\ 2}}V(S_{xy})$$
$$=\frac{1}{S_{xx}^{\ 2}}V\left(\sum(x_i-\bar{x})(y_i-\bar{y})\right)$$
$$=\frac{1}{S_{xx}^{\ 2}}V\left(\sum(x_i-\bar{x})y_i-\sum(x_i-\bar{x})(\bar{y})\right)$$
$$=\frac{1}{S_{xx}^{\ 2}}V\left(\sum(x_i-\bar{x})y_i-\bar{y}\sum(x_i-\bar{x})\right)$$
$$=\frac{1}{S_{xx}^{\ 2}}V\left(\sum(x_i-\bar{x})y_i-\bar{y}\cdot0\right)$$
$$=\frac{1}{S_{xx}^{\ 2}}V\left(\sum(x_i-\bar{x})y_i\right)$$
$$=\frac{1}{S_{xx}^{\ 2}}V\left((x_1-\bar{x})y_1+\cdots+(x_n-\bar{x})y_n\right)$$

（確率変数は y_i だけだから）

$$=\frac{1}{S_{xx}^{\ 2}}\left((x_1-\bar{x})^2V(y_1)+\cdots+(x_n-\bar{x})^2V(y_n)\right)$$
$$=\frac{1}{S_{xx}^{\ 2}}\sum(x_i-\bar{x})^2V(y_i)$$

（ここで

$$V(y_i)=V(\beta_0+\beta_1x_i+\varepsilon_i)=V(\varepsilon_i)=\sigma^2$$

となるから）

$$=\frac{1}{S_{xx}^2}\sum(x_i-\bar{x})^2\sigma^2$$
$$=\frac{\sigma^2}{S_{xx}^2}\sum(x_i-\bar{x})^2=\frac{\sigma^2}{S_{xx}}$$

したがって、

$$\boxed{V(b_1)=\frac{\sigma^2}{S_{xx}}}$$

（2）標本の切片の期待値と分散

1）切片 b_0 の期待値

$$E(b_0)=E(\bar{y}-b_1\bar{x})$$
$$=E(\bar{y})-E(b_1\bar{x})$$
$$=E(\bar{y})-\bar{x}E(b_1)$$
$$=E(\beta_0+\beta_1\bar{x})-\bar{x}\beta_1$$
$$=E(\beta_0)+E(\beta_1\bar{x})-\bar{x}\beta_1$$
$$=\beta_0+\beta_1\bar{x}-\beta_1\bar{x}=\beta_0$$

したがって、

$$\boxed{E(b_0)=\beta_0}$$

2）切片 b_0 の分散

$$
\begin{aligned}
&V\left(b_0\right)=V\left(\bar{y}-b_1\bar{x}\right)\\
&=V\left(\frac{1}{n}\sum y_i-\frac{S_{xy}}{S_{xx}}\bar{x}\right)\\
&=V\left(\frac{1}{n}\sum y_i-\frac{\sum\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{S_{xx}}\bar{x}\right)\\
&=V\left(\frac{1}{n}\sum y_i-\frac{\sum\left(x_i-\bar{x}\right)y_i}{S_{xx}}\bar{x}\right)\\
&=V\left(\frac{1}{n}\sum y_i-\sum\frac{\left(x_i-\bar{x}\right)\bar{x}}{S_{xx}}y_i\right)\\
&=V\left(\sum\frac{1}{n}y_i-\sum\frac{\left(x_i-\bar{x}\right)\bar{x}}{S_{xx}}y_i\right)\\
&=V\left(\sum\left\{\frac{1}{n}-\frac{\left(x_i-\bar{x}\right)\bar{x}}{S_{xx}}\right\}y_i\right)\\
&=\sum\left(\frac{1}{n}-\frac{\left(x_i-\bar{x}\right)\bar{x}}{S_{xx}}\right)^2V\left(y_i\right)\\
&=\sigma^2\sum\left(\frac{1}{n}-\frac{\left(x_i-\bar{x}\right)\bar{x}}{S_{xx}}\right)^2\\
&=\sigma^2\sum\left(\frac{1}{n^2}-\frac{2\left(x_i-\bar{x}\right)\bar{x}}{nS_{xx}}+\frac{\left(x_i-\bar{x}\right)^2\bar{x}^2}{{S_{xx}}^2}\right)\\
&=\sigma^2\left(\frac{1}{n}+\sum\frac{2\left(x_i-\bar{x}\right)\bar{x}}{nS_{xx}}+\sum\frac{\left(x_i-\bar{x}\right)^2\bar{x}^2}{{S_{xx}}^2}\right)\\
&=\sigma^2\left(\frac{1}{n}+\frac{2\bar{x}}{nS_{xx}}\sum\left(x_i-\bar{x}\right)+\frac{\bar{x}^2}{{S_{xx}}^2}\sum\left(x_i-\bar{x}\right)^2\right)\\
&=\sigma^2\left(\frac{1}{n}-\frac{0}{nS_{xx}}+\frac{\bar{x}^2S_{xx}}{{S_{xx}}^2}\right)=\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)
\end{aligned}
$$

これはさらに次のようになる。

$$
\begin{aligned}
&V\left(b_0\right)=\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{\mathrm{S}_{xx}}\right)=\sigma^2\left(\frac{\mathrm{S}_{xx}+n\bar{x}^2}{n\mathrm{S}_{xx}}\right)\\
&=\sigma^2\left(\frac{\sum\left(x_i-\bar{x}\right)^2+n\bar{x}^2}{n\mathrm{S}_{xx}}\right)\\
&=\sigma^2\left(\frac{\left\{\sum x_i^2-2\bar{x}\sum x_i+n\bar{x}^2\right\}+n\bar{x}^2}{n\mathrm{S}_{xx}}\right)\\
&=\sigma^2\left(\frac{\left\{\sum x_i^2-2n\bar{x}^2+n\bar{x}^2\right\}+n\bar{x}^2}{n\mathrm{S}_{xx}}\right)\\
&=\frac{\sigma^2\sum {x_i}^2}{nS_{xx}}
\end{aligned}
$$

したがって、

$$
\boxed{V\left(b_0\right)=\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{\mathrm{S}_{xx}}\right)=\frac{\sigma^2\sum {x_i}^2}{nS_{xx}}}
$$

（3）標本の切片と回帰係数の共分散

切片 b_0 と回帰係数 b_1 の共分散がどうなるかも明らかにしておこう。

一般に次の式が成り立つ。

$$
Cov(X,Y)=E\left(XY\right)-E\left(X\right)E\left(Y\right)
$$

これを使う。

$$
\begin{aligned}
&Cov\left(b_0,b_1\right)\\
&=Cov\left(\bar{y}-\frac{S_{xy}}{S_{xx}}\bar{x},\quad\frac{S_{xy}}{S_{xx}}\right)\\
&=Cov\left(\begin{array}{l}\sum\frac{1}{n}y_i-\sum\frac{\left(x_i-\bar{x}\right)\bar{x}}{S_{xx}}y_i,\\ \frac{\sum\left(x_j-\bar{x}\right)\left(y_j-\bar{y}\right)}{S_{xx}}\end{array}\right)\\
&=Cov\left(\sum_i\left\{\frac{1}{n}-\frac{\left(x_i-\bar{x}\right)\bar{x}}{S_{xx}}\right\}y_i,\sum_j\frac{\left(x_j-\bar{x}\right)}{S_{xx}}y_j\right)
\end{aligned}
$$

$$= E\left(\sum_i \left\{ \frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}} \right\} y_i \sum_j \frac{(x_j - \bar{x})}{S_{xx}} y_j \right)$$

$$-E\left(\sum_i \left\{ \frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}} \right\} y_i \right) E\left(\sum_j \frac{(x_j - \bar{x})}{S_{xx}} y_j \right)$$

$$= E\left(\sum_i \sum_j \left\{ \frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}} \right\} y_i \frac{(x_j - \bar{x})}{S_{xx}} y_j \right)$$

$$-E\left(\sum_i \left\{ \frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}} \right\} y_i \right) E\left(\sum_j \frac{(x_j - \bar{x})}{S_{xx}} y_j \right)$$

（確率変数は y_i だけだから）

$$= E\left(\sum_i \sum_j \left\{ \frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}} \right\} \left(\frac{(x_j - \bar{x})}{S_{xx}} \right) y_i y_j \right)$$

$$-\sum_i \left\{ \frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}} \right\} E(y_i)$$

$$\times \sum_j \left(\frac{(x_j - \bar{x})}{S_{xx}} \right) E(y_j)$$

$$= \sum_i \sum_j \left\{ \frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}} \right\} \left(\frac{(x_j - \bar{x})}{S_{xx}} \right) \times E(y_i y_j)$$

$$-\sum_i \sum_j \left\{ \frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}} \right\} \left(\frac{(x_j - \bar{x})}{S_{xx}} \right)$$

$$\times E(y_i) E(y_j)$$

$$= \sum_i \sum_j \left\{ \frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}} \right\} \left(\frac{(x_j - \bar{x})}{S_{xx}} \right)$$

$$\times \left(E(y_i y_j) - E(y_i) E(y_j) \right)$$

$$= \sum_i \sum_j \left\{ \frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}} \right\} \left(\frac{(x_j - \bar{x})}{S_{xx}} \right)$$

$$\times Cov(y_i, y_j)$$

（i≠j のとき $Cov(y_i, y_j) = 0$ となるので、二重のシグマは不要だから）

$$= \sum_i \left\{ \frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}} \right\} \left(\frac{(x_i - \bar{x})}{S_{xx}} \right) Cov(y_i, y_i)$$

（i のとき

$$Cov(y_i, y_i) = V(y_i)$$

$$= V(\beta_0 + \beta_1 x_i + \varepsilon_i) = V(\varepsilon_i) = \sigma^2$$

となるから）

$$= \sum_i \left\{ \frac{1}{n} \left(\frac{(x_i - \bar{x})}{S_{xx}} \right) - \frac{(x_i - \bar{x})^2 \bar{x}}{{S_{xx}}^2} \right\} V(y_i)$$

$$= \sigma^2 \left\{ \frac{1}{n} \sum \left(\frac{(x_i - \bar{x})}{S_{xx}} \right) - \frac{\bar{x}}{{S_{xx}}^2} \sum (x_i - \bar{x})^2 \right\}$$

$$= \sigma^2 \left\{ -\frac{\bar{x}}{{S_{xx}}^2} \sum (x_i - \bar{x})^2 \right\} = \sigma^2 \left\{ -\frac{\bar{x}}{{S_{xx}}^2} S_{xx} \right\}$$

$$= -\frac{\sigma^2 \bar{x}}{S_{xx}}$$

したがって、

$$\boxed{Cov(b_0, b_1) = -\frac{\sigma^2 \bar{x}}{S_{xx}}}$$

（4）標本の残差分散と母集団の誤差分散

これらをもとに、標本の残差分散の期待値 $E(V_e)$ は母集団の誤差 ε の分散 σ^2 であることを示す。明らかにしたいのは次の式だ。

$$E(V_e) = E\left(\frac{\sum {e_i}^2}{n-2} \right) = \sigma^2$$

まず、母集団の誤差について

$$\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$$

$$= (y_i) - \beta_0 - \beta_1 x_1$$

$$= (b_0 + b_1 x_i + e_i) - \beta_0 - \beta_1 x_i$$

$$= (b_0 - \beta_0) + (b_1 - \beta_1) x_i + e_i$$

誤差の 2 乗は、

$$\begin{aligned}\varepsilon_i^2 &= \left((b_0-\beta_0)+(b_1-\beta_1)x_i+e_i\right)^2\\ &=(b_0-\beta_0)^2+(b_1-\beta_1)^2x_i^2+e_i^2\\ &\quad+2(b_0-\beta_0)(b_1-\beta_1)x_i\\ &\quad+2(b_0-\beta_0)e_i+2(b_1-\beta_1)x_ie_i\end{aligned}$$

誤差の平方和は、

$$\begin{aligned}&\sum\varepsilon_i^2\\ &=n(b_0-\beta_0)^2+(b_1-\beta_1)^2\sum x_i^2+\sum e_i^2\\ &\quad+2(b_0-\beta_0)(b_1-\beta_1)\sum x_i\\ &\quad+2(b_0-\beta_0)\sum e_i\\ &\quad+2(b_1-\beta_1)\sum x_ie_i\\ &=n(b_0-\beta_0)^2+(b_1-\beta_1)^2\sum x_i^2+\sum e_i^2\\ &\quad+2n(b_0-\beta_0)(b_1-\beta_1)\bar{x}\\ &=\sum e_i^2+n(b_0-\beta_0)^2+(b_1-\beta_1)^2\sum x_i^2\\ &\quad+2n(b_0-\beta_0)(b_1-\beta_1)\bar{x}\end{aligned}$$

この式は後でも使うことになる。A という名前をつけておこう。

$$\begin{aligned}A:\sum\varepsilon_i^2&=\sum e_i^2\\ &\quad+n(b_0-\beta_0)^2\\ &\quad+(b_1-\beta_1)^2\sum x_i^2\\ &\quad+2n(b_0-\beta_0)(b_1-\beta_1)\bar{x}\end{aligned}$$

誤差の平方和の期待値は、

$$\begin{aligned}&E\left(\sum\varepsilon_i^2\right)\\ &=E\left(\begin{array}{l}\sum e_i^2+n(b_0-\beta_0)^2+(b_1-\beta_1)^2\sum x_i^2\\ +2n(b_0-\beta_0)(b_1-\beta_1)\bar{x}\end{array}\right)\\ &=E\left(\sum e_i^2\right)+nE(b_0-\beta_0)^2\\ &\quad+E(b_1-\beta_1)^2\sum x_i^2\\ &\quad+2nE\left((b_0-\beta_0)(b_1-\beta_1)\bar{x}\right)\\ &=E\left(\sum e_i^2\right)+nE\left(b_0-E(b_0)\right)^2\\ &\quad+E\left(b_1-E(b_1)\right)^2\sum x_i^2\\ &\quad+2nE\left(\left(b_0-E(b_0)\right)\left(b_1-E(b_1)\right)\bar{x}\right)\\ &=E\left(\sum e_i^2\right)+nV(b_0)+V(b_1)\sum x_i^2\\ &\quad+2nCov(b_0,b_1)\bar{x}\end{aligned}$$

一般に、確率変数を X とするとき、分散と期待値の関係は次の式で表されるが、

$$V(X)=E\left(X-E(X)\right)^2$$

これを利用すると、

$$V(\varepsilon_i)=E(\varepsilon_i-0)^2=E(\varepsilon_i^2)$$

また、回帰分析の仮定より、

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

$$V(\varepsilon_i)=\sigma^2$$

したがって、

$$\boxed{E(\varepsilon_i^2)=\sigma^2}$$

ここから

$$\begin{aligned}E\left(\sum\varepsilon_i^2\right)&=E\left(\underbrace{\sigma^2+\sigma^2+\cdots+\sigma^2}_{n}\right)\\ &=E(n\sigma^2)=n\sigma^2\end{aligned}$$

また、上で見てきたことより

$$\boxed{Cov(b_0,b_1)=-\frac{\sigma^2\bar{x}}{S_{xx}}}$$

$$\boxed{V(b_0)=\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)=\frac{\sigma^2\sum x_i^2}{nS_{xx}}}$$

$$\boxed{V(b_1)=\frac{\sigma^2}{S_{xx}}}$$

以上をもとに、

$$\begin{aligned}
&E\left(\sum \varepsilon_i^{\,2}\right)=n\sigma^2\\
&=E\left(\sum e_i^{\,2}\right)+nV(b_0)+V(b_1)\sum x_i^{\,2}\\
&\quad+2nCov(b_0,b_1)\bar{x}\\
&=E\left(\sum e_i^{\,2}\right)+n\left(\frac{\sigma^2\sum x_i^{\,2}}{nS_{xx}}\right)+\left(\frac{\sigma^2}{S_{xx}}\right)\sum x_i^{\,2}\\
&\quad-2n\left(\frac{\sigma^2\bar{x}}{S_{xx}}\right)\bar{x}\\
&=E\left(\sum e_i^{\,2}\right)+\frac{\sigma^2\sum x_i^{\,2}}{S_{xx}}+\frac{\sigma^2\sum x_i^{\,2}}{S_{xx}}-\frac{2n\sigma^2\bar{x}^2}{S_{xx}}\\
&=E\left(\sum e_i^{\,2}\right)+\frac{2\sigma^2\sum x_i^{\,2}-2\sigma^2n\bar{x}^2}{S_{xx}}\\
&=E\left(\sum e_i^{\,2}\right)+\frac{2\sigma^2}{S_{xx}}\left(\sum x_i^{\,2}-n\bar{x}^2\right)\\
&=E\left(\sum e_i^{\,2}\right)+\frac{2\sigma^2}{S_{xx}}\sum\left(x_i-\bar{x}\right)^2\\
&=E\left(\sum e_i^{\,2}\right)+\frac{2\sigma^2}{S_{xx}}S_{xx}\\
&=E\left(\sum e_i^{\,2}\right)+2\sigma^2
\end{aligned}$$

ここから、

$$\begin{aligned}
&n\sigma^2=E\left(\sum e_i^{\,2}\right)+2\sigma^2\\
&E\left(\sum e_i^{\,2}\right)=(n-2)\sigma^2\\
&E\left(\frac{\sum e_i^{\,2}}{n-2}\right)=E(V_e)=\sigma^2
\end{aligned}$$

したがって、

$$\boxed{E(V_e)=E\left(\frac{\sum e_i^{\,2}}{n-2}\right)=E\left(\frac{S_e}{n-2}\right)=\sigma^2}$$

6.4 回帰係数・切片・残差平方和の確率分布

（1）確率分布の相互関係

上では、標本統計量の期待値と分散がどうなるのを示したが、どのような分布に従うのかはまだ明らかにしていない。ここではそのあたりのことを解説するが、その前にまず代表的な確率分布の相互関係についておさらいしておこう。

Z を標準正規分布に従う確率変数とし、K_n を自由度 n の χ^2 分布に従う確率変数とする。また、T_n を自由度 n の t 分布に従う確率変数とし、F_n^m を自由度 m,n の F 分布に従う確率変数とする。このとき次の式が成り立つ（以下の式の右辺の確率変数がすべて独立であるとする）。

$$\begin{aligned}
&K_1=Z^2\\
&K_n=Z_1^2+Z_2^2+\cdots+Z_n^2\\
&T_n=\frac{Z}{\sqrt{K_n/n}}\\
&F_n^m=\frac{K_m/m}{K_n/n}
\end{aligned}$$

以下では次のことが重要になる。

$$\begin{aligned}
&T_{n-2}=\frac{Z}{\sqrt{K_{n-2}/(n-2)}}\\
&\left(T_{n-2}\right)^2=\frac{Z^2}{K_{n-2}/(n-2)}=\frac{K_1/1}{K_{n-2}/(n-2)}=F_{n-2}^1
\end{aligned}$$

上の式は、自由度 n−2 の t 分布がどういうものになるのかを、正規分布と χ^2 分布に関係させて示したものである。下の式は、自由度 n−2 の t 分布に従う変数の 2 乗は自由度（1,n−2）の F 分布に従うということを意味している。

（2）直交行列と直交変換

さて、標本回帰係数がなぜ t 分布に従うのかを明らかにするためには、直交行列というものについて知っておく必要がある。直交行列とは転置行列と逆行列が等しくなる正方行列のことだ。

直交行列には次の性質がある。直交行列の各列ベクトルは大きさ 1 で互いに直交する。直交行列の各行ベクトルは大きさ 1 で互いに直交する。直交行列の逆行列は直交行列である。以下のものは直交行列の例だ。

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \quad \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$$

直交行列 T を用いてベクトルを別のベクトルに変換できる。これを直交変換という。

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = T \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

さて、$x_1, x_2, \cdots, x_n$ が独立でどれも $N\left(0, \sigma^2\right)$ に従う確率変数とし、それを並べてベクトルとするとき、これを直交変換したベクトルの要素である $w_1, w_2, \cdots, w_n$ も独立でどれも $N\left(0, \sigma^2\right)$ に従うことになる。この証明には別の知識が必要なのでここでは前提と考えてほしい。

$$\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} = T \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

（3）回帰係数と切片の確率分布

1）変換に用いる直交行列

ここで次のような直行行列 T を考える。3 行目以降はこの行列を直交行列にするような任意の要素が入っているものとする[4]。

$$T = \begin{bmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \frac{x_1 - \bar{x}}{\sqrt{S_{xx}}} & \frac{x_2 - \bar{x}}{\sqrt{S_{xx}}} & \cdots & \frac{x_n - \bar{x}}{\sqrt{S_{xx}}} \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \end{bmatrix}$$

ここで、y_i についての母集団のそれぞれの誤差を $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ として、次のような直交変換を考える。

$$\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} = T \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

このとき、仮定より、$\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ は独立でどれも $N\left(0, \sigma^2\right)$ に従うから、$w_1, w_2, \cdots, w_n$ も独立で、どれも $N\left(0, \sigma^2\right)$ に従うことになる。

[4] 鈴木・山田（2004: 270-273）、小寺（1986:97-99）等を参照。

この直交変換を行うと、

$$w_1 = \frac{1}{\sqrt{n}}\varepsilon_1 + \frac{1}{\sqrt{n}}\varepsilon_2 + \cdots + \frac{1}{\sqrt{n}}\varepsilon_n$$

$$= \frac{1}{\sqrt{n}}\sum \varepsilon_i = \frac{1}{\sqrt{n}}\left(n\bar{\varepsilon}\right) = \sqrt{n}\bar{\varepsilon}$$

したがって、

$$\boxed{\sqrt{n}\bar{\varepsilon} = w_1 \sim N\left(0, \sigma^2\right)}$$

ここで、

$$\left[\begin{array}{l} \bar{y} = \beta_0 + \beta_1\bar{x} + \bar{\varepsilon} \\ \quad \therefore \bar{\varepsilon} = \bar{y} - \beta_0 - \beta_1\bar{x} \end{array}\right.$$

だから、

$$\boxed{\sqrt{n}\left(\bar{y} - \beta_0 - \beta_1\bar{x}\right) = w_1 \sim N\left(0, \sigma^2\right)}$$

次にw_2のほうの直交変換について。

$$w_2 = \frac{x_1 - \bar{x}}{\sqrt{S_{xx}}}\varepsilon_1 + \frac{x_2 - \bar{x}}{\sqrt{S_{xx}}}\varepsilon_2 + \cdots + \frac{x_n - \bar{x}}{\sqrt{S_{xx}}}\varepsilon_n$$

$$= \sum \frac{x_i - \bar{x}}{\sqrt{S_{xx}}}\varepsilon_i = \frac{1}{\sqrt{S_{xx}}}\sum \left(x_i - \bar{x}\right)\varepsilon_i$$

$$= \frac{1}{\sqrt{S_{xx}}}\left(\sum \left(x_i - \bar{x}\right)\left\{y_i - \left(\beta_0 + \beta_1 x_i\right)\right\}\right)$$

$$= \frac{1}{\sqrt{S_{xx}}}\left(\sum \left(x_i - \bar{x}\right)\left(y_i - \beta_0 - \beta_1 x_i\right)\right)$$

$$= \frac{1}{\sqrt{S_{xx}}}\left(\begin{array}{l} \sum \left(x_i - \bar{x}\right)y_i - \beta_0\sum \left(x_i - \bar{x}\right) \\ -\beta_1\sum \left(x_i - \bar{x}\right)x_i \end{array}\right)$$

$$= \frac{1}{\sqrt{S_{xx}}}\left(\sum \left(x_i - \bar{x}\right)y_i - 0 - \beta_1\sum \left(x_i - \bar{x}\right)^2\right)$$

（ここで

$$\left[\begin{array}{l} b_1 = \dfrac{S_{xy}}{S_{xx}} = \dfrac{\sum \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum \left(x_i - \bar{x}\right)^2} \\ = \dfrac{\sum \left(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y}\right)}{\sum \left(x_i - \bar{x}\right)^2} \\ = \dfrac{\sum x_i y_i - n\bar{x}\bar{y} - \sum \bar{x}y_i + n\bar{x}\bar{y}}{\sum \left(x_i - \bar{x}\right)^2} \\ = \dfrac{\sum x_i y_i - \sum \bar{x}y_i}{\sum \left(x_i - \bar{x}\right)^2} = \dfrac{\sum \left(x_i - \bar{x}\right)y_i}{\sum \left(x_i - \bar{x}\right)^2} \\ \therefore \sum \left(x_i - \bar{x}\right)y_i = b_1\sum \left(x_i - \bar{x}\right)^2 \end{array}\right.$$

だから）

$$= \frac{1}{\sqrt{S_{xx}}}\left(b_1\sum \left(x_i - \bar{x}\right)^2 - \beta_1\sum \left(x_i - \bar{x}\right)^2\right)$$

$$= \frac{1}{\sqrt{S_{xx}}}\left(b_1 S_{xx} - \beta_1 S_{xx}\right) = \frac{S_{xx}}{\sqrt{S_{xx}}}\left(b_1 - \beta_1\right)$$

$$= \sqrt{S_{xx}}\left(b_1 - \beta_1\right)$$

したがって、

$$\boxed{\sqrt{S_{xx}}\left(b_1 - \beta_1\right) = w_2 \sim N\left(0, \sigma^2\right)}$$

2）回帰係数と切片の確率分布

すぐ上の式より、

$$\sqrt{S_{xx}}\left(b_1 - \beta_1\right) = w_2 \sim N\left(0, \sigma^2\right)$$

$$\left(b_1 - \beta_1\right) = \frac{w_2}{\sqrt{S_{xx}}} \sim N\left(0, \frac{\sigma^2}{S_{xx}}\right)$$

$$b_1 = \frac{w_2}{\sqrt{S_{xx}}} + \beta_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

したがって、

$$\boxed{b_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)}$$

すなわち、b_1 の平均は β_1、分散は σ^2/S_{xx} であり（これらは 6.3 示したものと同じ）、それは正規分布に従う。

また、ここから次のこともいえる。

$$\frac{b_1-\beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1)$$

次に b_0 について。正規分布に従う次の式があった。

$$\sqrt{n}\left(\bar{y}-\beta_0+\beta_1\bar{x}\right)=w_1 \sim N\left(0,\sigma^2\right)$$

これを変形する。

$$\sqrt{n}\left(\bar{y}-\beta_0-\beta_1\bar{x}\right)=w_1$$

$$\bar{y}=\beta_0+\beta_1\bar{x}+\frac{w_1}{\sqrt{n}}$$

ところで、$\bar{y}$ については次の式が成り立つ。

$$\bar{y}=b_0+b_1\bar{x}$$

2 つ上の式にこれを代入すると次の式が導ける。

$$b_0+b_1\bar{x}=\beta_0+\beta_1\bar{x}+\frac{w_1}{\sqrt{n}}$$

$$b_0=\beta_0+\beta_1\bar{x}+\frac{w_1}{\sqrt{n}}-b_1\bar{x}$$

ここで右辺の確率変数は w_1 と b_1 だが、これらはともに独立に正規分布に従う。したがって左辺の b_0 も正規分布に従うことになる。

b_0 の期待値と分散はすでに上で求めているが、ここでの考え方からも簡単に導ける。まず、期待値は次のようになる。

$$E\left(b_0\right)=E\left(\beta_0+\beta_1\bar{x}+\frac{w_1}{\sqrt{n}}-b_1\bar{x}\right)$$

$$=\beta_0+\beta_1\bar{x}+\frac{1}{\sqrt{n}}E\left(w_1\right)-E\left(b_1\right)\bar{x}$$

$$=\beta_0+\beta_1\bar{x}+\frac{1}{\sqrt{n}}(0)-\beta_1\bar{x}=\beta_0$$

また b_0 の分散は、次のようになる。

$$V\left(b_0\right)=V\left(\beta_0+\beta_1\bar{x}+\frac{w_1}{\sqrt{n}}-b_1\bar{x}\right)$$

$$=V\left(\frac{w_1}{\sqrt{n}}-b_1\bar{x}\right)=\frac{1}{n}V\left(w_1\right)+\bar{x}^2V\left(b_1\right)$$

$$=\frac{\sigma^2}{n}+\bar{x}^2\frac{\sigma^2}{S_{xx}}=\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)$$

したがって、b_0 は次の正規分布に従う。

$$b_0 \sim N\left(\beta_0,\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)\right)$$

また、ここから次のこともいえる。

$$\frac{b_0-\beta_0}{\sqrt{\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}} \sim N(0,1)$$

(4) 残差の平方和と χ^2 分布

ところで、母集団の誤差の平方和と標本の残差の平方和については次の A の式が成り立った。

$$A:\sum \varepsilon_i^{\,2}=\sum e_i^{\,2}+n\left(b_0-\beta_0\right)^2+\left(b_1-\beta_1\right)^2\sum x_i^{\,2}+2n\left(b_0-\beta_0\right)\left(b_1-\beta_1\right)\bar{x}$$

これは上で用いた 2 つの式の 2 乗（＝下の C 式の 2 乗）を用いた B の式と同等である（証明は後で示す）。

$$B:\sum \varepsilon_i^{\,2}=\sum e_i^{\,2}+n\left(\bar{y}-\beta_0-\beta_1\bar{x}\right)^2+S_{xx}\left(b_1-\beta_1\right)^2$$

$$C:\begin{cases}\sqrt{n}\left(\bar{y}-\beta_0-\beta_1\bar{x}\right)=w_1\sim N\left(0,\sigma^2\right)\\ \sqrt{S_{xx}}\left(b_1-\beta_1\right)=w_2\sim N\left(0,\sigma^2\right)\end{cases}$$

この B 式を使って、残差平方和と χ^2 分布の関係について考えていく。

まず、次の上下の分布が等しいことを確認しておこう。

$$\begin{cases}\sum w_i^2=w_1^2+w_2^2+\cdots+w_n^2\\ \sum \varepsilon_i^2=\varepsilon_1^2+\varepsilon_2^2+\cdots+\varepsilon_n^2\end{cases}$$

$\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$ はどれも独立で $N\left(0,\sigma^2\right)$ に従い、$w_1,w_2,\cdots,w_n$ もどれも独立で $N\left(0,\sigma^2\right)$ に従う。それゆえ、それぞれの平方和である $\sum \varepsilon_i^2$ の分布と $\sum w_i^2$ の分布は等しくなるのである。

さて、前項での解説（＝C 式）から B 式は次のように書けることをわれわれは知っている。

$$\begin{aligned}&\sum \varepsilon_i^{\,2}\\ &=\sum e_i^{\,2}+n\left(\bar{y}-\beta_0-\beta_1\bar{x}\right)^2+S_{xx}\left(b_1-\beta_1\right)^2\\ &=\sum e_i^{\,2}+w_1^2+w_2^2\end{aligned}$$

これを利用すると、上のペアの式は、次のように表現できる。

$$\begin{cases}w_1^2+w_2^2+\cdots+w_n^2\\ \sum e_i^{\,2}+w_1^2+w_2^2\end{cases}$$

このペアの上下の式は同じ分布になり、そこから、次のペアの上下の式も同じ分布になる。

$$\begin{cases}w_3^2+w_4^2+\cdots+w_n^2\\ \sum_{i=1}^{n} e_i^2\end{cases}$$

上の式の添え字が 3 から始まり、下の式の添え字が 1 から始まることに注意しよう。

このペアの上の式を σ^2 で割ると次のようになる。

$$\left(\frac{w_3}{\sigma}\right)^2+\left(\frac{w_4}{\sigma}\right)^2+\cdots+\left(\frac{w_n}{\sigma}\right)^2$$

これは標準正規分布に従う n−2 個の確率変数の 2 乗和であり、自由度 n−2 の χ^2 分布に従うことになる。

したがって、下の式を σ^2 で割ると、これも自由度 n−2 の χ^2 分布に従うことになる。

$$\begin{cases} \dfrac{1}{\sigma^2}\left(w_3^2+w_4^2+\cdots+w_n^2\right) \sim \chi_{n-2}^2 \\ \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n} e_i^2 \sim \chi_{n-2}^2 \end{cases}$$

以上から、次のことがわかる。

$$\boxed{\frac{1}{\sigma^2}\sum_{i=1}^{n} e_i^2 = \frac{S_e}{\sigma^2} \sim \chi_{n-2}^2}$$

この「χ^2分布に従う、誤差の分散σ^2と残差の平方和S_eの式」はとても重要な式である。

ところで、標本の残差の分散は次の式で定義されていた。

$$V_e = \frac{S_e}{n-2} = \frac{\sum e_i^2}{n-2}$$

この定義によると、次のものが自由度 n−2 のχ^2分布に従うことになる。

$$\boxed{\frac{(n-2)V_e}{\sigma^2} \sim \chi_{n-2}^2}$$

この「χ^2分布に従う、誤差の分散σ^2と残差の分散V_eの式」が後で利用されることになる。

(5) 検定統計量と t 分布

1) 回帰係数

以上をもとに、回帰係数の検定統計量の確率分布がなぜ自由度 n−2 の t 分布になるのかを示そう。回帰係数の検定統計量は次のものだった。

$$Test = \frac{b_1-\beta_1}{\sqrt{\dfrac{V_e}{S_{xx}}}}$$

これは次のように変形できる。

$$Test = \frac{b_1-\beta_1}{\sqrt{\dfrac{V_e}{S_{xx}}}} = \frac{\left(b_1-\beta_1\right)\Big/\sqrt{\dfrac{\sigma^2}{(n-2)}}}{\sqrt{\dfrac{V_e}{S_{xx}}}\Big/\sqrt{\dfrac{\sigma^2}{(n-2)}}}$$

$$= \frac{\dfrac{\left(b_1-\beta_1\right)\sqrt{(n-2)}}{\sqrt{\sigma^2}}}{\sqrt{\dfrac{(n-2)V_e}{S_{xx}\sigma^2}}} = \frac{\dfrac{\left(b_1-\beta_1\right)\sqrt{(n-2)S_{xx}}}{\sqrt{\sigma^2}}}{\sqrt{\dfrac{(n-2)V_e}{\sigma^2}}}$$

$$= \frac{\dfrac{\left(b_1-\beta_1\right)\sqrt{S_{xx}}}{\sqrt{\sigma^2}}}{\sqrt{\dfrac{(n-2)V_e}{\sigma^2}\Big/(n-2)}}$$

$$= \frac{\dfrac{b_1-\beta_1}{\sqrt{\sigma^2/S_{xx}}}}{\sqrt{\dfrac{(n-2)V_e}{\sigma^2}\Big/(n-2)}} \qquad (*最後の式)$$

ここで、分子ならびに分母の中身が次の確率分布に関連することが見て取れる。

$$\frac{b_1-\beta_1}{\sqrt{\dfrac{\sigma^2}{S_{xx}}}} \sim N(0,1)$$

$$\frac{(n-2)V_e}{\sigma^2} \sim \chi_{n-2}^2$$

2 つの独立な確率変数 Z と K_n があり、Z は標準正規分布 N(0,1)に、K_n は自由度 n のχ^2分布に

従うときの $X=\dfrac{Z}{\sqrt{K_n/n}}$ の従う分布が自由度 n の t 分布である[5]。

さきの式の変形の「＊最後の式」は下の形になっている。

$$X=\frac{Z}{\sqrt{K_{n-2}/(n-2)}}$$

したがって、結局、下の統計量は自由度 n−2 の t 分布に従うことになる。

$$\boxed{\frac{b_1-\beta_1}{\sqrt{\dfrac{V_e}{S_{xx}}}}\sim t_{n-2}}$$

また、$\beta_1=0$ という帰無仮説を検定する場合は、次の検定統計量が自由度 n−2 の t 分布に従うことを利用する。

$$Test\left(H_0:\beta_1=0\right)=\frac{b_1}{\sqrt{\dfrac{V_e}{S_{xx}}}}\sim t_{n-2}$$

2）切片

切片の検定統計量は下の式だった。

$$Test=\frac{b_0-\beta_0}{\sqrt{V_e\left(\dfrac{1}{n}+\dfrac{\bar{x}^2}{S_{xx}}\right)}}$$

これは次のように変形できる。

$$\begin{aligned}
Test&=\frac{b_0-\beta_0}{\sqrt{V_e\left(\dfrac{1}{n}+\dfrac{\bar{x}^2}{S_{xx}}\right)}}\\
&=\frac{(b_0-\beta_0)\Big/\sqrt{\dfrac{\sigma^2}{(n-2)}}}{\sqrt{V_e\left(\dfrac{1}{n}+\dfrac{\bar{x}^2}{S_{xx}}\right)}\Big/\sqrt{\dfrac{\sigma^2}{(n-2)}}}\\
&=\frac{(b_0-\beta_0)\Big/\sqrt{\dfrac{1}{(n-2)}}\sqrt{\sigma^2\left(\dfrac{1}{n}+\dfrac{\bar{x}^2}{S_{xx}}\right)}}{\sqrt{V_e}\Big/\sqrt{\dfrac{\sigma^2}{(n-2)}}}\\
&=\frac{(b_0-\beta_0)\Big/\sqrt{\sigma^2\left(\dfrac{1}{n}+\dfrac{\bar{x}^2}{S_{xx}}\right)}}{\sqrt{\dfrac{V_e}{(n-2)}}\Big/\sqrt{\dfrac{\sigma^2}{(n-2)}}}\\
&=\frac{(b_0-\beta_0)}{\sqrt{\dfrac{\sigma^2\left(\dfrac{1}{n}+\dfrac{\bar{x}^2}{S_{xx}}\right)}{\dfrac{V_e(n-2)}{(n-2)\sigma^2}}}}\\
&=\frac{b_0-\beta_0}{\sqrt{\dfrac{\sigma^2\left(\dfrac{1}{n}+\dfrac{\bar{x}^2}{S_{xx}}\right)}{\dfrac{(n-2)V_e}{\sigma^2}\Big/(n-2)}}}
\end{aligned}$$

（＊最後の式）

ここで、分子ならびに分母の中身が次の確率分布に関連することが見て取れる。

[5] t 分布を考案したゴセット（スチューデント）は、分子と分母に σ を置いて、σ 自体を消してしまおうと考えたようだ（小寺, 1986: 101）。そうなれば検定に母集団の σ は不要になる。なるほど冴えてるなあ。

$$\frac{b_0-\beta_0}{\sqrt{\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}}\sim N(0,1)$$

$$\frac{(n-2)V_e}{\sigma^2}\sim \chi^2_{n-2}$$

2 つの独立な確率変数 Z と K_n があり、Z は標準正規分布 N(0,1)に、K_n は自由度 n の χ^2 分布に従うときの $X=\frac{Z}{\sqrt{K_n/n}}$ の従う分布が自由度 n の t 分布である。

さきの式の変形の「＊最後の式」は下の形になっている。

$$X=\frac{Z}{\sqrt{K_{n-2}/(n-2)}}$$

だから、結局下の統計量は自由度 n−2 の t 分布に従うことになる。

$$\boxed{\frac{b_0-\beta_0}{\sqrt{V_e\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}}\sim t_{n-2}}$$

また、$\beta_0=0$ という帰無仮説を検定する場合は、次の検定統計量が自由度 n−2 の t 分布に従うことを利用する。

$$Test(H_0:\beta_0=0)=\frac{b_0}{\sqrt{V_e\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}}\sim t_{n-2}$$

(6) 回帰係数と切片の信頼区間

自由度 n−2 の t 分布の上側確率 2.5%に対応する値を $t_{n-2}(0.025)$ で表すとき、回帰係数の 95%の信頼区間は次のように導ける。

$$P\left(\begin{array}{l}-t_{n-2}(0.025)\le\dfrac{b_1-\beta_1}{\sqrt{\dfrac{V_e}{S_{xx}}}}\\ \le t_{n-2}(0.025)\end{array}\right)=0.95$$

$$\Leftrightarrow P\left(\begin{array}{l}-t_{n-2}(0.025)\sqrt{\dfrac{V_e}{S_{xx}}}\le b_1-\beta_1\\ \le t_{n-2}(0.025)\sqrt{\dfrac{V_e}{S_{xx}}}\end{array}\right)=0.95$$

$$\Leftrightarrow P\left(\begin{array}{l}-b_1-t_{n-2}(0.025)\sqrt{\dfrac{V_e}{S_{xx}}}\le-\beta_1\\ \le -b_1+t_{n-2}(0.025)\sqrt{\dfrac{V_e}{S_{xx}}}\end{array}\right)=0.95$$

$$\Leftrightarrow P\left(\begin{array}{l}b_1+t_{n-2}(0.025)\sqrt{\dfrac{V_e}{S_{xx}}}\ge\beta_1\\ \ge b_1-t_{n-2}(0.025)\sqrt{\dfrac{V_e}{S_{xx}}}\end{array}\right)=0.95$$

$$\Leftrightarrow P\left(\begin{array}{l}b_1-t_{n-2}(0.025)\sqrt{\dfrac{V_e}{S_{xx}}}\le\beta_1\\ \le b_1+t_{n-2}(0.025)\sqrt{\dfrac{V_e}{S_{xx}}}\end{array}\right)=0.95$$

切片の 95%の信頼区間は次のように導ける。

$$P\left(\begin{array}{l}-t_{n-2}(0.025)\le\dfrac{b_0-\beta_0}{\sqrt{V_e\left(\frac{1}{n}+\frac{\bar{x}^2}{S_x}\right)}}\\ \le t_{n-2}(0.025)\end{array}\right)=0.95$$

$$\Leftrightarrow P\left(\begin{array}{l}-t_{n-2}(0.025)\sqrt{V_e\left(\frac{1}{n}+\frac{\bar{x}^2}{S_x}\right)}\\ \le b_0-\beta_0\\ \le t_{n-2}(0.025)\sqrt{V_e\left(\frac{1}{n}+\frac{\bar{x}^2}{S_x}\right)}\end{array}\right)=0.95$$

$$\Leftrightarrow P\left(\begin{array}{l}-b_0-t_{n-2}(0.025)\sqrt{V_e\left(\frac{1}{n}+\frac{\bar{x}^2}{S_x}\right)}\\ \le -\beta_0\\ \le -b_0+t_{n-2}(0.025)\sqrt{V_e\left(\frac{1}{n}+\frac{\bar{x}^2}{S_x}\right)}\end{array}\right)=0.95$$

$$\Leftrightarrow P\left(\begin{array}{l}b_0+t_{n-2}(0.025)\sqrt{V_e\left(\frac{1}{n}+\frac{\bar{x}^2}{S_x}\right)}\\ \ge \beta_0\\ \ge b_0-t_{n-2}(0.025)\sqrt{V_e\left(\frac{1}{n}+\frac{\bar{x}^2}{S_x}\right)}\end{array}\right)=0.95$$

$$\Leftrightarrow P\left(\begin{array}{l}b_0-t_{n-2}(0.025)\sqrt{V_e\left(\frac{1}{n}+\frac{\bar{x}^2}{S_x}\right)}\\ \le \beta_0\\ \le b_0+t_{n-2}(0.025)\sqrt{V_e\left(\frac{1}{n}+\frac{\bar{x}^2}{S_x}\right)}\end{array}\right)=0.95$$

（7）回帰係数の検定と相関係数の検定

ここで、回帰係数（傾き）の検定（$\beta_1=0$）と相関係数の検定（$\rho=0$）の検定における統計量についてみておこう。結論からいうと、どちらも自由度 n－2 の t 分布に従う 2 つの検定統計量は同じものだ。すなわち、次の 3 つの式が成り立つ。

$$Test(H_0:\rho=0)=\frac{r}{\sqrt{(1-r^2)/(n-2)}}\sim t_{n-2}$$

$$Test(H_0:\beta_1=0)=\frac{b_1}{\sqrt{\frac{V_e}{S_{xx}}}}\sim t_{n-2}$$

$$\frac{r}{\sqrt{(1-r^2)/(n-2)}}=\frac{b_1}{\sqrt{\frac{V_e}{S_{xx}}}}$$

このことを明らかにしておこう。

$$Test(H_0:\beta_1=0)=\frac{b_1}{\sqrt{\frac{V_e}{S_{xx}}}}=\frac{\frac{S_{xy}}{S_{xx}}}{\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{S_{xx}(n-2)}}}$$

$$=\frac{\frac{S_{xy}\sqrt{S_{xx}}\sqrt{n-2}}{S_{xx}}}{\sqrt{\sum\left(y_i-\{b_1(x_i-\bar{x})+\bar{y}\}\right)^2}}$$

$$=\frac{\frac{S_{xy}\sqrt{n-2}}{\sqrt{S_{xx}}}}{\sqrt{\sum(y_i-b_1(x_i-\bar{x})-\bar{y})^2}}$$

$$=\frac{\frac{S_{xy}\sqrt{n-2}}{\sqrt{S_{xx}}}}{\sqrt{\sum\left((y_i-\bar{y})-b_1(x_i-\bar{x})\right)^2}}$$

$$=\frac{\frac{S_{xy}\sqrt{n-2}}{\sqrt{S_{xx}}}}{\sqrt{\sum\left(\begin{array}{l}(y_i-\bar{y})^2-2b_1(x_i-\bar{x})(y_i-\bar{y})\\ +b_1^2(x_i-\bar{x})^2\end{array}\right)}}$$

$$=\frac{\frac{S_{xy}\sqrt{n-2}}{\sqrt{S_{xx}}}}{\sqrt{\begin{array}{l}\sum(y_i-\bar{y})^2-2b_1\sum(x_i-\bar{x})(y_i-\bar{y})\\ +b_1^2\sum(x_i-\bar{x})^2\end{array}}}$$

$$
= \frac{\dfrac{S_{xy}\sqrt{n-2}}{\sqrt{S_{xx}}}}{\sqrt{S_{yy}-2b_1S_{xy}+b_1^2S_{xx}}}
$$

$$
= \frac{\dfrac{S_{xy}\sqrt{n-2}}{\sqrt{S_{xx}}}}{\sqrt{S_{yy}-2\left(\dfrac{S_{xy}}{S_{xx}}\right)S_{xy}+\left(\dfrac{S_{xy}}{S_{xx}}\right)^2S_{xx}}}
$$

$$
= \frac{\dfrac{S_{xy}\sqrt{n-2}}{\sqrt{S_{xx}}}}{\sqrt{S_{yy}-\dfrac{2S_{xy}{}^2}{S_{xx}}+\dfrac{S_{xy}{}^2}{S_{xx}}}} = \frac{\dfrac{S_{xy}\sqrt{n-2}}{\sqrt{S_{xx}}}}{\sqrt{\dfrac{S_{xx}S_{yy}-S_{xy}{}^2}{S_{xx}}}}
$$

$$
= \frac{S_{xy}\sqrt{n-2}}{\sqrt{S_{xx}S_{yy}-S_{xy}{}^2}} = \frac{S_{xy}\sqrt{n-2}}{\sqrt{S_{xx}S_{yy}}\sqrt{1-\dfrac{S_{xy}{}^2}{S_{xx}S_{yy}}}}
$$

$$
= \frac{\dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}\left(\sqrt{n-2}\right)}{\sqrt{1-\dfrac{S_{xy}{}^2}{S_{xx}S_{yy}}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}
$$

$$
= \frac{r}{\sqrt{\left(1-r^2\right)/\left(n-2\right)}} = Test\left(H_0:\rho=0\right)
$$

（8）A 式＝B 式の証明

6.4 （4）で解説を省略した A と B の式が同じであることを示しておく。

$$
\begin{aligned}
A:\sum\varepsilon_i{}^2 &= \sum e_i{}^2 \\
&+ n\left(b_0-\beta_0\right)^2 \\
&+ \left(b_1-\beta_1\right)^2\sum x_i{}^2 \\
&+ 2n\left(b_0-\beta_0\right)\left(b_1-\beta_1\right)\bar{x}
\end{aligned}
$$

$$
\begin{aligned}
B:\sum\varepsilon_i{}^2 &= \sum e_i{}^2 \\
&+ n\left(\bar{y}-\beta_0-\beta_1\bar{x}\right)^2 \\
&+ S_{xx}\left(b_1-\beta_1\right)^2
\end{aligned}
$$

このとき、

$$
\bar{y} = b_0 + b_1\bar{x}
$$

$$
S_{xx} = \sum\left(x_i-\bar{x}\right)^2 = \sum x_i^2 - n\bar{x}^2
$$

であるから、B 式は

$$
\begin{aligned}
B:\sum\varepsilon_i{}^2 &= \sum e_i{}^2 \\
&+ n\left(b_0+b_1\bar{x}-\beta_0-\beta_1\bar{x}\right)^2 \\
&+ \left(b_1-\beta_1\right)^2\left(\sum x_i^2-n\bar{x}^2\right) \\
&= \sum e_i{}^2 + n\left\{\left(b_0-\beta_0\right)+\left(b_1-\beta_1\right)\bar{x}\right\}^2 \\
&+ \left(b_1-\beta_1\right)^2\left(\sum x_i^2-n\bar{x}^2\right)
\end{aligned}
$$

とできる。

$$
b_0-\beta_0 = P
$$

$$
b_1-\beta_1 = Q
$$

とおくならば、A 式、B 式はそれぞれ

$$
A:\sum e_i{}^2 + nP^2 + Q^2\sum x_i{}^2 + 2nPQ\bar{x}
$$

$$
B:\sum e_i{}^2 + n\left(P+Q\bar{x}\right)^2 + Q^2\left(\sum x_i^2-n\bar{x}^2\right)
$$

と表せる。

ここで、A—B をすると、

$$
\begin{aligned}
A-B &= \left(\sum e_i{}^2 + nP^2 + Q^2\sum x_i{}^2 + 2nPQ\bar{x}\right) \\
&\quad -\left(\sum e_i{}^2 + n\left(P+Q\bar{x}\right)^2 + Q^2\left(\sum x_i^2-n\bar{x}^2\right)\right) \\
&= nP^2 + Q^2\sum x_i{}^2 + 2nPQ\bar{x} \\
&\quad - n\left(P^2+2PQ\bar{x}+Q^2\bar{x}^2\right) - Q^2\left(\sum x_i^2-n\bar{x}^2\right) \\
&= nP^2 + Q^2\sum x_i{}^2 + 2nPQ\bar{x} - nP^2 - 2nPQ\bar{x} \\
&\quad - nQ^2\bar{x}^2 - Q^2\sum x_i^2 + nQ^2\bar{x}^2 = 0
\end{aligned}
$$

したがって、A 式と B 式は等しい。

6.5 回帰モデルの検討

（1）母回帰係数$\beta_1=0$のt検定とF検定

回帰分析においては平方和や分散をもとに、回帰式の表す回帰モデルに意味があるか、回帰モデルは何らかの説明力を持っているのかどうかを検討することがある。ここではこの検討について解説するが、その前に帰無仮説$\beta_1=0$のもうひとつの検定法について述べておく必要がある。

帰無仮説$\beta_1=0$の検定の検定統計量は次のようなものだった。

$$Test\left(H_0:\beta_1=0\right)=\frac{b_1}{\sqrt{V_e/\mathrm{S}_{xx}}}\sim t_{n-2}$$

自由度nのt分布に従う変数の2乗は自由度(1,n)のF分布に従うので、このt分布を用いた検定は、次のF分布を用いた検定と同等である。

$$Test\left(H_0:\beta_1=0\right)=\frac{b_1^2}{V_e/\mathrm{S}_{xx}}\sim F_{n-2}^1$$

（2）3つの平方和に関する等式

1）回帰モデルを平方和から検討する等式

平方和や分散をもとに回帰モデルの有効性を検討する際に、検討の基礎となる式は全平方和、回帰平方和、残差平方和に関する次の式である。

$$S_{yy}=S_e+S_R$$

その中身は次のようになっている。

$$\sum\left(y_i-\bar{y}\right)^2=\sum\left(y_i-\hat{y}_i\right)^2+\sum\left(\hat{y}_i-\bar{y}\right)^2$$

すなわち、「全平方和＝残差平方和＋回帰平方和」なのである。

2）等式が成り立つわけ

この等式が成り立つわけは次の通りだ。

$$\begin{aligned}
S_{yy}&=\sum\left(y_i-\bar{y}\right)^2\\
&=\sum\left\{y_i-\left(b_0+b_1x_i\right)+\left(b_0+b_1x_i\right)-\bar{y}\right\}^2\\
&=\sum\left\{\left(y_i-\left(b_0+b_1x_i\right)\right)+\left(\left(b_0+b_1x_i\right)-\bar{y}\right)\right\}^2\\
&=\sum\left\{y_i-\left(b_0+b_1x_i\right)\right\}^2\\
&\quad+\sum\left\{\left(b_0+b_1x_i\right)-\bar{y}\right\}^2\\
&\quad+2\sum\left\{y_i-\left(b_0+b_1x_i\right)\right\}\left\{\left(b_0+b_1x_i\right)-\bar{y}\right\}\\
&=S_e+S_R+2\sum e_i\left\{\left(b_0+b_1x_i\right)-\bar{y}\right\}\\
&=S_e+S_R+2\sum\left\{\left(b_0+b_1x_i\right)e_i-\bar{y}e_i\right\}\\
&=S_e+S_R+2\left(\sum\left\{b_0e_i+b_1x_ie_i-\bar{y}e_i\right\}\right)\\
&=S_e+S_R+2\left(\sum\left\{b_0e_i-\bar{y}e_i+b_1x_ie_i\right\}\right)\\
&=S_e+S_R+2\left(\sum\left\{\left(b_0-\bar{y}\right)e_i+b_1x_ie_i\right\}\right)\\
&=S_e+S_R+2\left(\left(b_0-\bar{y}\right)\sum e_i+b_1\sum x_ie_i\right)
\end{aligned}$$

（ここで、最小二乗法の条件より

$\sum e_i=0,\quad \sum x_ie_i=0$

だから）

$$\begin{aligned}
&=S_e+S_R+2\left(\left(b_0-\bar{y}\right)\cdot0+b_1\cdot0\right)\\
&=S_e+S_R
\end{aligned}$$

したがって、

$$\boxed{S_{yy}=S_e+S_R}$$

（3）回帰分析における諸検定の相互関係

1）R^2の検討

回帰分析では、平方和や分散をもとに回帰式を検討するとき、決定係数R^2が用いられることがある。決定係数は、上の回帰平方和を全平方和で割

ったものだ。それは、標本の全変動のどの程度が回帰変動で説明されるかを意味しており、寄与率と呼ばれることもある。

$$R^2 = \frac{S_R}{S_{yy}} = 1 - \frac{S_e}{S_{yy}}$$

説明変数が 1 つである回帰分析の場合、決定係数は相関係数の 2 乗となる。そうなる理由はこうである。

まず、回帰平方和は次のようになる。

$$S_R = \sum \left(\hat{y}_i - \bar{y}\right)^2 = \sum \left(b_0 + b_1 x_i - \left(b_0 + b_1 \bar{x}\right)\right)^2$$
$$= \sum \left(b_1 \left(x_i - \bar{x}\right)\right)^2 = b_1^2 \sum \left(x_i - \bar{x}\right)^2$$
$$= \left(\frac{S_{xy}}{S_{xx}}\right)^2 S_{xx} = \frac{S_{xy}^2}{S_{xx}}$$

これを決定係数の式に代入すると、次のようになり、決定係数が相関係数の 2 乗になることがわかる。

$$R^2 = \frac{S_R}{S_{yy}} = \frac{S_{xy}^2 / S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$
$$= \left(\frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}\right)^2 = r^2$$

相関係数 r の検定は次の検定統計量で行われる。

$$Test\left(H_0 : \rho = 0\right) = \frac{r}{\sqrt{\left(1 - r^2\right) / \left(n - 2\right)}} \sim t_{n-1}$$

したがって、決定係数 R^2 は上の検定量の 2 乗である次の式によって検討されることになる。t 分布に従う変数を 2 乗しているので、これは F 分布に従うことになる。

$$\boxed{\frac{R^2}{\left(1 - R^2\right) / \left(n - 2\right)} \sim F_{n-2}^1}$$

このテストで有意であれば、回帰モデルは説明力を持ち、意味のあるものとされる。

2）分散比の検討

回帰モデルが説明力をもつか、回帰に意味があるかということは次の式でも検討される。

$$\boxed{\frac{S_R / 1}{S_e / \left(n - 2\right)} = \frac{S_R}{V_e} \sim F_{n-2}^1}$$

この式の分子は回帰の平方和をその自由度 1 で割ったもの、分母は残差平方和をその自由度 n−2 で割ったものである。このように平方和をその自由度で割ったものを平均平方という。この式は回帰の平均平方を残差の平均平方で割ったものだが、これを分散比という。分散比が大きいとき、この回帰モデルは意味のあるものとされる。

ここで、この式を少し変形していこう。上でも使った、

$$S_R = \frac{S_{xy}^2}{S_{xx}}$$

を用いる。

$$\frac{S_R / 1}{S_e / \left(n - 2\right)} = \frac{S_R}{V_e} = \frac{S_{xy}^2 / S_{xx}}{V_e} = \frac{S_{xy}^2 / S_{xx}^2}{V_e / \mathrm{S}_{xx}}$$
$$= \frac{\left(S_{xy} / S_{xx}\right)^2}{V_e / \mathrm{S}_{xx}} = \frac{{b_1}^2}{V_e / \mathrm{S}_{xx}}$$
$$= test\left(\beta_1 = 0\right)$$

ここからわかることは、この分散比の検討の式は

この項の最初（6.5 （1））で紹介した回帰係数のF検定の式と同じということだ。したがって、分散比を用いた回帰式の検討では、母回帰係数$\beta_1 = 0$という帰無仮説が検定されていることになる。

分散比の式は次のようにも変形できる。

$$\frac{S_R/1}{S_e/(n-2)} = \frac{\frac{S_R}{S_{yy}}}{\frac{S_e}{S_{yy}}\Big/(n-2)}$$

$$= \frac{\frac{S_R}{S_{yy}}}{\left(\frac{S_{yy}-S_R}{S_{yy}}\right)\Big/(n-2)} = \frac{\frac{S_R}{S_{yy}}}{\left(1-\frac{S_R}{S_{yy}}\right)\Big/(n-2)}$$

$$= \frac{R^2}{(1-R^2)/(n-2)}$$

この最後の式は決定係数の検討に用いた式（相関係数の検定統計量の2乗の式）である。

これまで見てきたことから言えることは次のことだ。説明変数が1つの単回帰分析において、回帰係数の検定、分散比の検討、決定係数の検討はすべて同等のものである。それらはすべて母回帰係数$\beta_1 = 0$という帰無仮説を検定していると言っていい。そしてそれら検定は母相関係数$\rho = 0$の検定と同等なものとも言える。

6.6　課題

（1）問題

a）y_iの平均とyの平均はどう違うのかを述べよ。

b）yの平均について、母集団の回帰係数と切片、誤差の平均、xの平均を使って表せ。

c）μの平均について、母集団の回帰係数と切片、xの平均を使って表せ。

d）yの平均について、標本の回帰係数と切片、xの平均を使って表せ。

e）$\hat{y}$の平均について、標本の回帰係数と切片、xの平均を使って表せ。

f）上のb～eの式で、確率変数となるのはどれか。

（2）解答

a）y_iはさまざまな値をとる1個の確率変数であり、その平均は期待値$E(y_i)$という意味である。他方、yの平均とは多数のy_iの平均を意味し、$\bar{y}$すなわち、$(\sum y_i)/n$のことである。母集団の分布が定まっているとき、前者は決まった値になるが、後者は値が確率的に変化する確率変数である。

b）

$$\bar{y} = \frac{1}{n}\sum y_i = \frac{1}{n}\sum(\beta_0 + \beta_1 x_i + \varepsilon_i)$$

$$= \frac{1}{n}\left(n\beta_0 + \beta_1\sum x_i + \sum\varepsilon_i\right) = \beta_0 + \beta_1\bar{x} + \bar{\varepsilon}$$

c）

$$\bar{\mu} = \frac{1}{n}\sum\mu_i = \frac{1}{n}\sum(\beta_0 + \beta_1 x_i)$$

$$= \frac{1}{n}\left(n\beta_0 + \beta_1\sum x_i\right) = \beta_0 + \beta_1\bar{x}$$

d）

$$\bar{y} = \frac{1}{n}\sum y_i = \frac{1}{n}\sum(b_0 + b_1 x_i + e_i)$$

$$= \frac{1}{n}\left(nb_0 + b_1\sum x_i + \sum e_i\right) = b_0 + b_1\bar{x}$$

e）

$$\bar{\hat{y}} = \frac{1}{n}\sum\hat{y}_i = \frac{1}{n}\sum(b_0 + b_1 x_i)$$

$$= \frac{1}{n}\left(nb_0 + b_1\sum x_i\right) = b_0 + b_1\bar{x}$$

f）確率変数になるのはb,d,e

7　おわりに

本セミナーでは 4 つの段階に分けて回帰分析について見てきた。3 節まで読んだ読者は回帰分析についての基本的な知識を得たと思う。4 節まで読んだ読者は回帰分析についての新しいイメージを得たのではないだろうか。5 節まで読んだ読者は、おおむね回帰分析の原理と検定や推定の方法の意味がわかっただろう。そして 6 節まで読んだ読者は、なぜ自由度 n－2 の t 分布や自由度（1,n－2）の F 分布が利用されるのかということが理解できただろう。

自身の研究に必要な理解の程度にはさまざまなものがあるとは思うが、深い知識を得るほど分析において自由になれるのは確実だと思う。ここがよくわからん、などと友人と話すのも楽しい。急ぐ必要はないので、ゆっくり考えればいい。

最後に、今回使ったデータを都道府県の自殺率のデータをもとにした重回帰分析の結果を載せておく（表 23、表 24）。ここでは説明変数が 1 つではなく複数である。今まで見てきた回帰分析の結果の表と見比べてほしい。回帰係数や平方和の表は単回帰分析の表にとても似ていることがわかるだろう。このような分析結果の意味を知るためにも、まず単回帰分析の原理を知っておく必要があるのだ。

表 23　重回帰分析の表 1（従属変数：自殺率）

	非標準化係数	標準誤差	標準化係数	t	有意確率
高齢化率	0.354	0.143	0.430	2.476	0.017
完全失業率	1.088	0.605	0.289	1.797	0.079
県民所得	0.052	0.739	0.012	0.071	0.944
［定数］	5.374	6.592		0.815	0.419
R	0.417				
R2乗	0.174				
調整済R2乗	0.116				

表 24　重回帰分析の表 2（従属変数：自殺率）

	平方和	自由度	平均平方	F	有意確率
回帰	41.271	3	13.757	3.009	0.040
残差	196.565	43	4.571		
合計	237.836	46			

文献

Agresti, A. & B. Finlay, 1997, *Statistical Methods for the Social Sciences, (3rd ed.)*, Prentice Hall.

Bohrnstedt, G.W., & D. Knoke, 1994, *Statistics for Social Data Analysis, (3rd ed.)*, F. E. Peacock.

小寺平治, 1986『明解演習数理統計』共立出版.

小林久高, 2018a「母集団・標本・確率変数」『同志社社会学研究』22.

————, 2018b「離散型確率変数とその分布」『同志社社会学研究』22.

————, 2018c「連続型確率変数とその分布」『同志社社会学研究』22.

————, 2019a「統計的仮説検定の原理と実際」『同志社社会学研究』23.

————, 2019b「統計的推定の原理と実際」『同志社社会学研究』23.

蓑谷千凰彦, 2003『統計分布ハンドブック』朝倉書店.

永田靖・棟近雅彦, 2001『多変量解析法入門』サイエンス社.

鈴木武・山田朔太郎, 2004『数理統計学（改訂版）』内田老鶴圃.

安田三郎・原純輔, 1982『社会調査ハンドブック（第 3 版）』有斐閣.

資料

厚生労働省、2015「平成 27 年人口動態調査第 5.19 表　都道府県別にみた死因簡単分類別死亡率（人口 10 万対）」（2019 年 10 月 26 日取得，http://www.e-stat.go.jp/SG1/estat/）.

内閣府, 2017「平成 29 年版高齢社会白書（概要版）表 1-1-5　都道府県別高齢化率の推移」（2019 年 10 月 26 日取得，https://www8.cao.go.jp/kourei/whitepaper/w-2017/html/gaiyou）.

総務省統計局，2015「2015 年労働力調査（基本集計）第 6 表　完全失業率（モデル推計値、年平均）」（2019 年 10 月 26 日取得, https://www.stat.go.jp/data/roudou/pref/index.html）．
総務省統計局，2019「統計でみる都道府県のすがた 2019／社会生活統計指標　C．経済基盤」（2019 年 10 月 26 日取得，http://www.e-stat.go.jp/SG1/estat/）