慶應義塾大学ビジネス・スクール

回帰分析シリーズ１

— 単回帰分析 —

"... there is no one more dangerous than the unthinking user of a computer, who has no real understanding of the nature of the computations being processed inside the machine."

--- Johnston, J. "Econometric Method", 3rd.ed., 1984; pp.79. ---

単回帰分析は、すべての計量経済分析における基礎となるものである。したがってどのような複雑なモデルであっても、基本的には単回帰の延長線上にある。このノートでは、単回帰の概念や仮定、制約、計算方法を例題を用いて解説しているが、その目的は、回帰分析によって「何ができるか」を知ることだけでなく、その限界を理解することにある。

一般に、関係式に２つの変数しか含まれない確率モデル

$$Y_i = \alpha + \beta X_i + u_i \quad i = 1, 2, ..., n \tag{1}$$

は単純回帰と呼ばれる。Y は被説明変数あるいは従属変数、X は説明変数あるいは独立変数と呼ばれる。ここで α と β はそれぞれ、この関数の切片と勾配を表す未知のパラメータである。また u は Y の変動について X だけでは説明しきれない要因を表し、誤差項あるいは撹乱項と呼ばれる。

回帰モデルによって α と β を推定する場合、誤差項 u の確率分布にいくつかの仮定をおく必要がある。それは分布の平均値が０、分散が均一、自己相関しない、確率分布は正規分布にしたがうというものであり、それぞれ (2) から (4) のように表すことができる。

$$E(u_i) = 0 \qquad i = 1, 2, ..., n \tag{2}$$

$$E(u_i u_j) = \begin{cases} 0 & i \neq j \\ \sigma_u^2 & i = j \end{cases} \tag{3}$$

$$p(u_i) = N(0, \sigma^2) \tag{4}$$

これらの仮定は、一括して

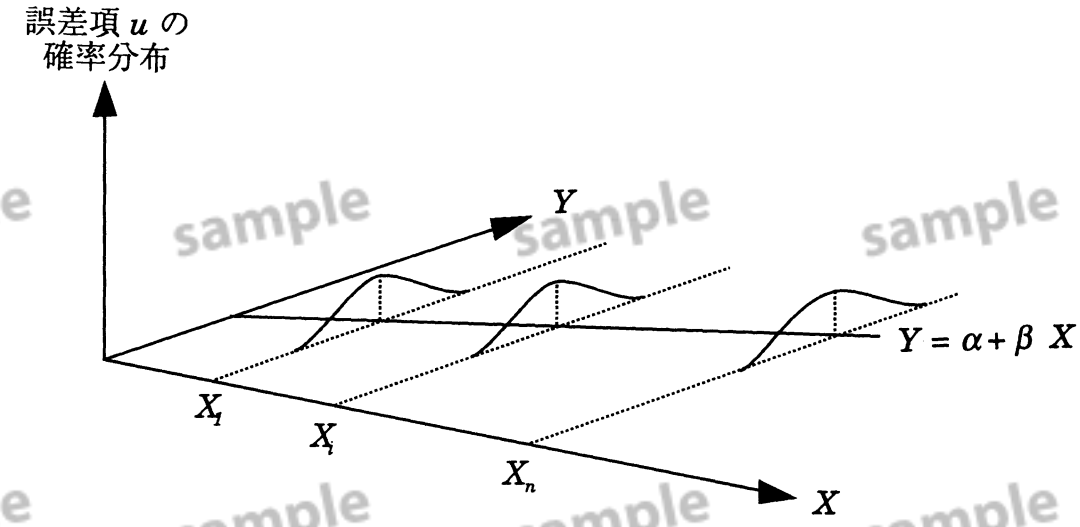
$$u = NID(0, \sigma_u^2)$$

によって示すこともできる。これは誤差項の平均が 0 で、正規に独立して分布してい

このノートは、慶應義塾大学ビジネス・スクールにおける補助教材として、同ビジネス・スクール教授矢作恒雄と博士課程磯辺剛彦が作成した。（1995年5月作成）

ることを意味する。特に分散の均一性については、図１に描かれているように、いずれのX_iにおいてもそのバラツキが同じであることを仮定する。

（図１）分散の均一性



最小二乗法

まずn個の観測データ、(X_1, Y_1)、(X_2, Y_2)、.... (X_n, Y_n) が図２のようにプロットされたとしよう。単純回帰の目的は、この直線を推定することにあり、$\hat{Y} = \hat{\alpha} + \hat{\beta} X$ で表されることにする。ここで、

$\hat{Y}$　：Xを固定したときの推定値Y

$\hat{\alpha}$、$\hat{\beta}$：パラメータの推定値（これ以降、$\hat{\alpha} = a$、$\hat{\beta} = b$とする）

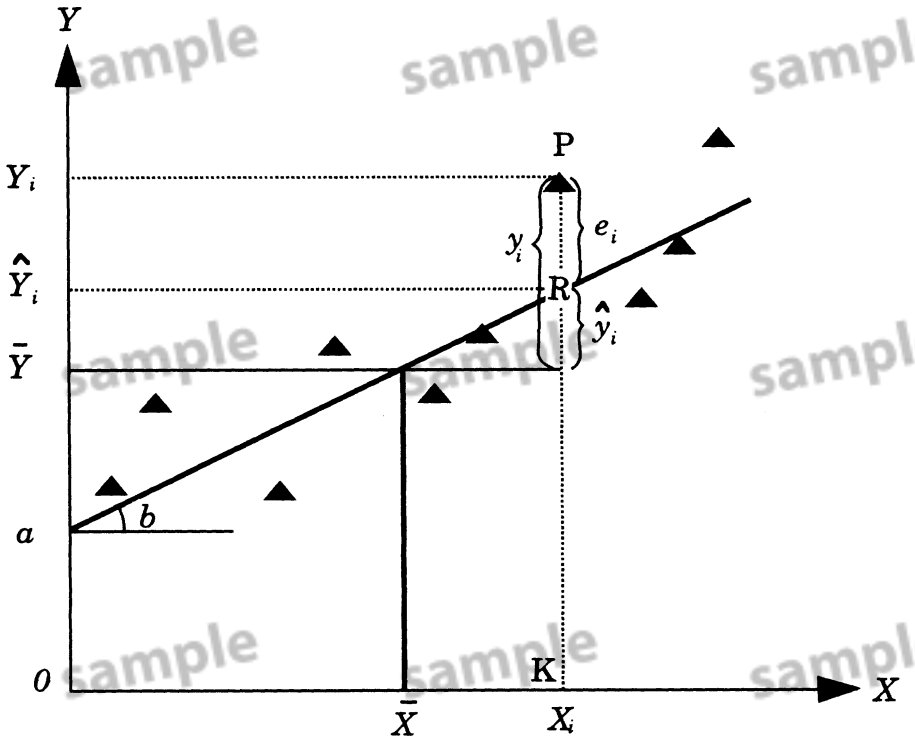
を示している。そして推定線を求めるには、ある公式を用いてaとbを観察データから導き出すことが必要になる。ここで任意の点P(X_i, Y_i)をとり、そのX軸への垂線を引く。垂線と推定線との交点をR$\left(X_i, \hat{Y}_i\right)$、X軸との交点をK$(X_i, 0)$とする。このときOK＝$X_i$、KP＝$Y_i$、KR＝$\hat{Y}_i$である。またX軸の垂線に沿った点Pと点Rとの差を

$$e_i = Y_i - \hat{Y}_i = Y_i - (a + b X_i) \tag{5}$$

と定義する。この推定線と観察値との差（残差）はプラスであったりマイナスであったりするが、これらの残差を二乗して和をとれば決してマイナスにはならない。最小二乗法とは、この残差平方和の値を最小にするようなaとbを選択するものである。そのための必要条件は、

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - a - b X_i)^2 \tag{6}$$

（図２）観測データと推定線



の最小化であり、(6) 式の残差平方和を a と b について偏微分したものが 0 となるような方程式を解けばよい[1)]。つまり、

$$\frac{\partial(\Sigma e^2)}{\partial a} = \frac{\partial(\Sigma e^2)}{\partial b} = 0 \tag{7}$$

の条件が必要になる。その結果、

$$\frac{\partial(\Sigma e^2)}{\partial a} = -2\,\Sigma\,(Y - a - b\,X) = 0$$
$$\Sigma\,Y = na + b\,\Sigma\,X \tag{8}$$

さらに、

$$\frac{\partial(\Sigma e^2)}{\partial b} = -2\,\Sigma\,X\,(Y - a - b\,X) = 0$$
$$\Sigma\,XY = a\,\Sigma\,X + b\,\Sigma\,X^2 \tag{9}$$

[1)] これ以降 $\sum_{i=1}^{n}$ を Σ に簡略化し、また各記号から i を省略する。たとえば、$\sum_{i=1}^{n} e_i$ は $\Sigma\,e$ と表す。

という２つの方程式が得られる。これらの方程式は正規方程式と呼ばれる。

このようにして推定された式は、データの平均点 $(\bar{X}, \bar{Y})$ を通るという特性をもつ。これは (8) 式を n で割ることによって証明される[1)] 。

$$\bar{Y} = a + b\,\bar{X} \tag{10}$$

さらに回帰係数 (regression coefficient) b は、$x = X - \bar{X}$、$y = Y - \bar{Y}$ という平均からの偏差を用いれば、

$$b = \Sigma xy \Big/ \Sigma x^2 \tag{11}$$

で示すことができる。これは (8) と (9) 式によって導き出される。

$$\begin{aligned} \Sigma XY &= a\,\Sigma X + b\,\Sigma X^2 \\ &= (\bar{Y} - b\,\bar{X})\,\Sigma X + b\,\Sigma X^2 \\ &= \frac{1}{n}(\Sigma Y)(\Sigma X) - \frac{b}{n}(\Sigma X)^2 + b\,\Sigma X^2 \end{aligned} \tag{12}$$

したがって、

$$\Sigma XY - \frac{1}{n}(\Sigma Y)(\Sigma X) = b\left[\Sigma X^2 - \frac{1}{n}(\Sigma X)^2\right] \tag{13}$$

ここで X と Y とをそれぞれの偏差 x と y とに置き換えれば、

$$\begin{aligned} \Sigma xy &= \Sigma (X - \bar{X})(Y - \bar{Y}) \\ &= \Sigma XY - \bar{Y}\,\Sigma X - \bar{X}\,\Sigma Y + n\,\bar{X}\,\bar{Y} \\ &= \Sigma XY - \frac{1}{n}(\Sigma X)(\Sigma Y) \end{aligned} \tag{14}$$

また、

$$\begin{aligned} \Sigma x^2 &= \Sigma (X - \bar{X})(X - \bar{X}) \\ &= \Sigma X^2 - \bar{X}\,\Sigma X - \bar{X}\,\Sigma X + n\,\bar{X}^2 \\ &= \Sigma X^2 - \frac{1}{n}(\Sigma X)^2 \end{aligned} \tag{15}$$

(14) 式と (15) 式とを (13) 式に代入すれば (11) 式が得られる。

また別解として、Σe^2 を最小化することによってもパラメータ b を求めることが

[1)] X と Y の算術平均は、それぞれ $\Sigma X / n$ と $\Sigma Y / n$ である。

可能である。(5) 式によりΣe^2は次のように変換される。

$$\begin{aligned}\Sigma e^2 &= \Sigma \left(Y - \hat{Y}\right)^2 \\ &= \Sigma \left[(Y - \bar{Y}) - \left(\hat{Y} - \bar{Y}\right)\right]^2 \\ &= \Sigma \left(y - \hat{y}\right)^2 \\ &= \Sigma \left(y - b\,x\right)^2\end{aligned} \tag{16}$$

ここで、

$$\hat{y} = b\,x \tag{17}$$

であることが分かる。なぜなら、

$$\begin{aligned}\hat{y} &= Y - \bar{Y} \\ &= a + b\,X - \left(a + b\,\bar{X}\right) \\ &= b\left(X - \bar{X}\right)\end{aligned} \tag{18}$$

Σe^2 を最小化するために、(16) 式を b について偏微分を行えば、

$$\frac{\partial(\Sigma e^2)}{\partial b} = 2\,\Sigma x\,(y - b\,x) = 0$$

したがって、

$$\Sigma x y = b\,\Sigma x^2$$

また切片のパラメータ a を求めるには、b を (10) 式に代入すればよい。

$$a = \bar{Y} - \left(\frac{\Sigma x y}{\Sigma x^2}\right)\bar{X} \tag{19}$$

平方和の分解

次に、総平方和 (TSS)、もしくは全変動と呼ばれるΣy^2は、説明された平方和 ($ESS : \Sigma \hat{y}^2$) と残差平方和 ($RSS : \Sigma e^2$) とに分解することができる。$y = \hat{y} + e$ は $y = b\,x + e$ に変換できるので、

$$\begin{aligned}\Sigma y^2 &= \Sigma (b x + e)^2 \\ &= b^2 \Sigma x^2 + \Sigma e^2 + 2 b e \Sigma x e \\ &= b^2 \Sigma x^2 + \Sigma e^2\end{aligned} \tag{20}$$

したがって、

$$\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS} \tag{21}$$

ここで、

$$\begin{aligned}\Sigma x e &= \Sigma \left(X - \bar{X} \right) e \\ &= \Sigma X e - \bar{X} \Sigma e \qquad \therefore \Sigma X e = \Sigma e = 0 \\ &= 0\end{aligned} \tag{22}$$

となる。これは (8) 式と (9) 式の偏微分が 0 であることを用いることによって簡単に証明できる。

$$\frac{\partial (\Sigma e^2)}{\partial a} = -2 \Sigma (Y - a - b X) = -2 \Sigma e = 0$$

さらに、

$$\frac{\partial (\Sigma e^2)}{\partial b} = -2 \Sigma X (Y - a - b X) = -2 \Sigma X e = 0$$

また ESS は (17) 式により b を消去し、x と y との関数で表すことができる。

$$\begin{aligned}\Sigma \hat{y}^2 &= \Sigma (b x)^2 \\ &= b^2 \Sigma x^2 \\ &= \left(\Sigma x y \Big/ \Sigma x^2 \right)^2 \Sigma x^2 \\ &= \left(\Sigma x y \right)^2 \Big/ \Sigma x^2\end{aligned} \tag{23}$$

したがって総平方和の分解することにより、回帰によって説明できる割合は、

$$\begin{aligned}\Sigma \hat{y}^2 \Big/ \Sigma y^2 &= \mathrm{ESS} \Big/ \mathrm{TSS} = (1 - \mathrm{RSS}) \Big/ \mathrm{TSS} \\ &= \left(\Sigma x y \right)^2 \Big/ \Sigma x^2 \Sigma y^2 \\ &= \left(\frac{\Sigma x y}{n} \right)^2 \Bigg/ \frac{\Sigma x^2}{n} \frac{\Sigma y^2}{n}\end{aligned} \tag{24}$$

通常これを決定係数と呼ぶ。決定係数はr_{xy}^2によって表され、0と1の間の数値になる。またr_{xy}は相関係数(correlation coefficient)と呼ばれる。相関係数とはXとYとの直線関係についての尺度であり、分子は共分散、分母はXとYの標準偏差の積で示される。

$$\begin{aligned} r &= \frac{\frac{\Sigma xy}{n}}{\sqrt{\frac{\Sigma x^2}{n}}\sqrt{\frac{\Sigma y^2}{n}}} \\ &= \frac{cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} \end{aligned} \tag{25}$$

最小二乗法の特性

最小二乗法が回帰分析にとって望ましい理由は、この方法によって推定されたパラメータが最良線形不偏推定値(best linear unbiased estimator ; BLUE)になりうるからである。そしてこの手法を、通常最小二乗法(Ordinary Least Squares : OLS)と呼ぶ。最良線形不偏推定値とは、主に次のような条件を満たすような推定値である。ここでは傾きbを例にすると、

(A) 推定されたbは線形関数であること

(B) bの平均値が真のパラメータβと等しいこと

(C) bの分布である$f(b)$の分散が他のいかなる推定値の分散よりも小さいこと

そして、この3つの条件を満たす図3の(b)が不偏推定値になる。

次に、(1)から(3)の仮定をおくことによって、未知数である切片aと傾きbの平均と分散を測定することができる。まず傾きbについて(11)式を用いて考察すると、

$$\begin{aligned} b &= \frac{\Sigma xy}{\Sigma x^2} \\ &= \frac{\Sigma x(Y-\bar{Y})}{\Sigma x^2} \\ &= \frac{\Sigma xY}{\Sigma x^2} \\ &= \Sigma wY \end{aligned} \tag{26}$$

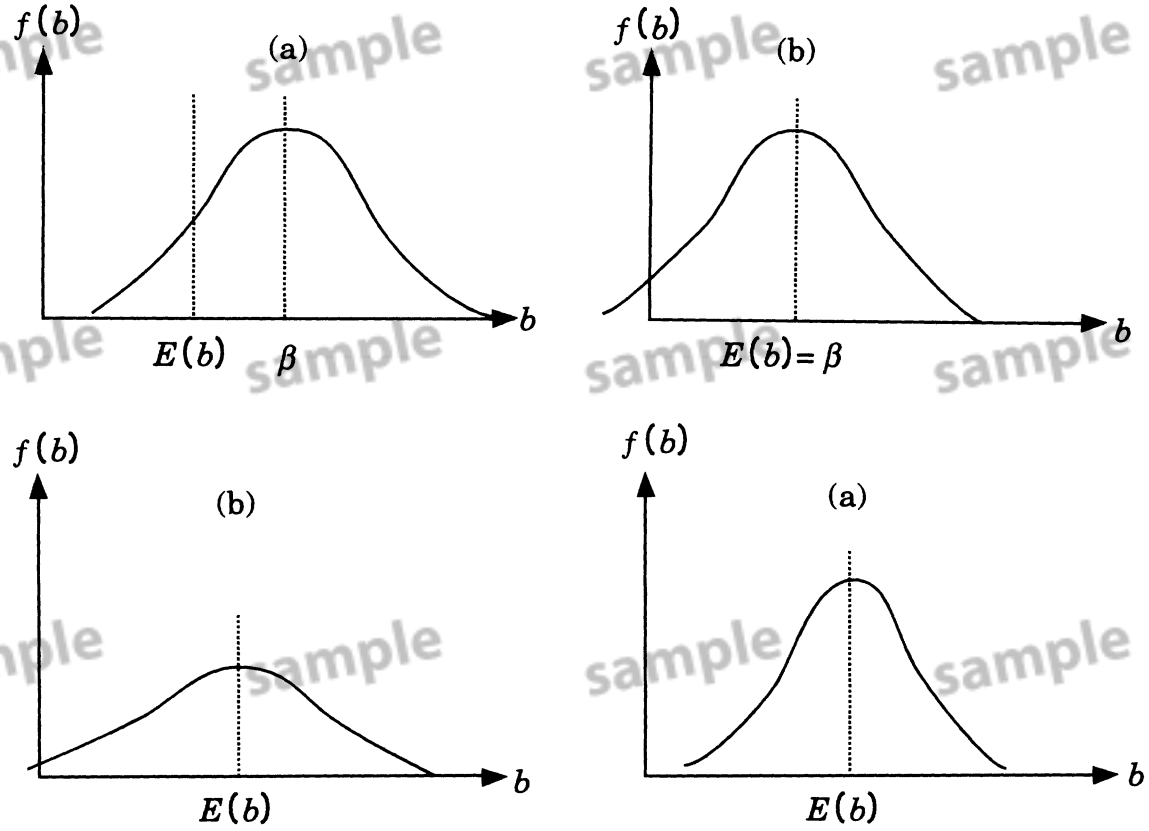
なぜなら、

$$\Sigma x\bar{Y} = \bar{Y}\Sigma(X-\bar{X}) = \bar{Y}(\Sigma X - \Sigma X) = 0 \tag{27}$$

ただし、

$$w = \frac{x}{\Sigma x^2} \tag{28}$$

（図３）不偏推定値の特性



この(28)式から、次のような３つの条件が導き出される。

$$\Sigma w = 0 \qquad \therefore \Sigma x = \Sigma \left(X - \bar{X}\right) = \Sigma X - \Sigma X \tag{29a}$$

$$\Sigma w X = 1 \qquad \therefore \Sigma w x = \Sigma x^2 \Big/ \Sigma x^2 = 1$$
$$= \Sigma w \left(X - \bar{X}\right) = \Sigma w X \tag{29b}$$

$$\Sigma w^2 = \Sigma x^2 \Big/ \left(\Sigma x^2\right)^2 = 1 \Big/ \Sigma x^2 \tag{29c}$$

これらの条件と(1)式を組み合わせることによって傾きbは、

$$b = \Sigma w Y = \Sigma w \left(\alpha + \beta X + u\right)$$
$$= \beta + \Sigma w u \tag{30}$$

そして(2)式の$E(u) = 0$の仮定を用いると、この(30)式の期待値は、

$$E(b) = \beta + E(\Sigma w u) = \beta \tag{31}$$

つまり最小二乗法によって求められた傾き b は、真の傾き β の不偏推定値であることがわかる。同様にして、傾き b の分散 (Var) については、

$$Var(b) = E[(b-\beta)^2] = E[(\Sigma w u)^2] \tag{32}$$

ここで、

$$\boldsymbol{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix}, \quad \boldsymbol{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

とすると、

$$\boldsymbol{w}'\boldsymbol{u} = w_1 u_1 + w_1 u_2 + \cdots + w_n u_{n-1} + w_n u_n$$

したがって、

$$(\Sigma w u)^2 = w_1^2 u_1^2 + \cdots + w_n^2 u_n^2 + 2 w_1 u_1 w_2 u_2 + \cdots + w_{n-1} u_n w_n u_{n-1}$$
$$= \Sigma w^2 u^2 + 2 \sum_{i<j} w_i u_i w_j u_j$$

ここで、(3) 式の仮定である $E(u^2) = \sigma_u^2$ と $E(u_i u_j) = 0$、さらに (29c) 式を用いれば、

$$Var(b) = E(\Sigma w u)^2 = \sigma_u^2 \Sigma w^2 = \sigma_u^2 / \Sigma x^2 \tag{33}$$

また切片 a については、(10) 式に (26) 式を代入することによって求めることができる。

$$\begin{aligned} a &= \bar{Y} - b\bar{X} \\ &= \bar{Y} - \bar{X}\Sigma w Y \qquad \therefore (26) \text{ 式} \\ &= \Sigma\left(\frac{1}{n} - w\bar{X}\right) Y \\ &= \Sigma\left(\frac{1}{n} - w\bar{X}\right)(\alpha + \beta X + u) \\ &= \alpha + \Sigma\left(\frac{1}{n} - w\bar{X}\right) u \end{aligned} \tag{34}$$

(34) 式の期待値をとると、

$$E(a)=\alpha \qquad \therefore E(u)=0 \tag{35}$$

(33) 式を用いると切片 a の分散は 、

$$\begin{aligned} Var(a) &= E\left[(a-\alpha)^2\right] \\ &= \Sigma\left(\frac{1}{n}-w\bar{X}\right)^2 u^2 \\ &= \sigma_u^2\left(\frac{1}{n}+\bar{X}^2\Sigma w^2-2\bar{X}\Sigma w\right) \\ &= \sigma_u^2\left(\frac{1}{n}+\frac{\bar{X}^2}{\Sigma x^2}\right) \end{aligned} \tag{36}$$

また傾き b と切片 a の共分散は 、

$$\begin{aligned} cov(a,b) &= E\left[(a-\alpha)(b-\beta)\right] \\ &= E\left\{\left[\bar{u}-(b-\beta)\bar{X}\right](b-\beta)\right\} \\ &= -\bar{X}\cdot E\left[(b-\beta)^2\right] \\ &= -\bar{X}\,\sigma_u^2 \Big/ \Sigma x^2 \qquad\qquad \therefore E(\bar{u})=0 \end{aligned} \tag{37}$$

以上のような最小二乗法による分散最小化は、仮定 (1) 式から (4) 式に基づいたものである。仮定 (1) 式は、a と b が u の線形関数であり、仮定 (4) 式は u が正規分布であることを示している。したがってa と b の平均と分散とが既知であれば、a と b は次のような正規分布にしたがうことになる。

$$a \sim N\left(\alpha, \sigma_u^2\left[\frac{1}{n}+\frac{\bar{X}^2}{\Sigma x^2}\right]\right) \tag{38}$$

$$b \sim N\left(\beta, \frac{\sigma_u^2}{\Sigma x^2}\right) \tag{39}$$

しかし u の分散であるσ_u^2は未知であるため、回帰分析を操作可能なものにするためにはσ_u^2の推定が必要になる。この推定には (5) 式を用いる。

$$\begin{aligned} e &= Y-\hat{Y} \\ &= (\alpha+\beta X+u)-(a+b X) \\ &= u-(a-\alpha)-(b-\beta)X \end{aligned} \tag{40}$$

さらに、$\bar{Y} = a + b\,\bar{X}$より、

$$\begin{aligned} a &= \left(\alpha + \beta\,\bar{X} + u\right) - b\,\bar{X} \\ &= \alpha - (b - \beta)\,\bar{X} + \bar{u} \end{aligned} \tag{41}$$

したがって、

$$e = (u - \bar{u}) - (b - \beta)\,x \tag{42}$$

ただし$\bar{u} \neq 0$ である。(42) 式の平方和は、

$$\begin{aligned} \Sigma e^2 &= \Sigma\,(u - \bar{u})^2 + (b - \beta)^2\,\Sigma x^2 - 2\,(b - \beta)\,\Sigma\,(u - \bar{u})\,x \\ &= \Sigma u^2 - n\,\bar{u}^2 + (b - \beta)^2\,\Sigma x^2 - 2\,(b - \beta)\,\Sigma\,u\,x \end{aligned} \tag{43}$$

残差平方和である (43) 式の期待値をとると、

$$\cdot\ E\left(\Sigma u^2\right) = n\,\sigma_u^2$$

$$\cdot\ E\left(\bar{u}^2\right) = Var\left(\bar{u}\right) = \sigma_u^2 \Big/ n$$

$$E\left[(b - \beta)^2\right] = Var\left(b\right) = \sigma_u^2 \Big/ \Sigma x^2$$

$$\begin{aligned} \cdot\ E\left[(b - \beta)\left(\Sigma\,u\,x\right)\right] &= E\left[\left(\Sigma\,w\,u\right)\left(\Sigma\,u\,x\right)\right] \\ &= \Sigma\,u\,x \Big/ \Sigma x^2 \cdot \Sigma\,u\,x \\ &= \sigma_u^2 \end{aligned}$$

より、

$$\begin{aligned} E\left(\Sigma e^2\right) &= n\,\sigma_u^2 - \sigma_u^2 + \sigma_u^2 - 2\,\sigma_u^2 \\ &= (n - 2)\,\sigma_u^2 \end{aligned} \tag{44}$$

つまり単回帰における誤差項の分散 σ_u^2 は、残差平方和をサンプル数から 2 を引いた数値によって割ったものになる。

$$E\left(\frac{\Sigma e^2}{n - 2}\right) = \sigma_u^2 \tag{45}$$

最小二乗法による推定

これまで見たように、仮定(1)式から(4)式のもとで、切片 a と傾き b の分布とを導き出してきた。さらに(38)式と(39)式とに含まれている未知の分散 σ_u^2 についても、その推定方法は既に述べたとおりである。次に考慮しなければならないのは、これらの仮定や帰結のもとで、どのように区間推定や仮説検定を行うべきかについてである。この推定・検定のプロセスについて解説する前に、さまざまな前提条件や命題が存在するため、これらについて一通り述べることにする。

定理1：確率変数 x の平均が μ であり、分散が σ^2 の正規分布であることが知られているとき、$z=\dfrac{x-\mu}{\sigma}$ は平均 0、分散 1 の正規分布にしたがう。

$$z=\frac{x-\mu}{\sigma} \sim N(0,1)$$

証明：x が確率変数であるとき、$z=\dfrac{x-\mu}{\sigma}$ もまた確率変数である。ここで $f(x)$ を x の確率密度関数とすると、z の平均は、

$$\mu_z=\int_{-\infty}^{+\infty}\frac{x-\mu}{\sigma}f(x)\,dx$$

$$=\frac{1}{\sigma}\left[\int_{-\infty}^{+\infty}x f(x)\,dx-\mu\int_{-\infty}^{+\infty}f(x)\,dx\right]$$

$$=\frac{1}{\sigma}(\mu-\mu)=0$$

また分散については、

$$\sigma_z^2=E[(z-\mu_z)^2]$$

$$=\int_{-\infty}^{+\infty}\left(\frac{x-\mu}{\sigma}\right)^2 f(x)\,dx-\mu_z^2$$

$$=\frac{1}{\sigma^2}\left[\int_{-\infty}^{+\infty}x^2 f(x)\,dx-2\mu\int_{-\infty}^{+\infty}x f(x)\,dx+\mu^2\int_{-\infty}^{+\infty}f(x)\,dx\right] \quad \therefore \mu_z=0$$

$$=\frac{1}{\sigma^2}[(\sigma^2+\mu^2)-2\mu\cdot\mu+\mu^2]=1$$

定理2：$z_1, z_2, \cdots, z_n$ が互いに独立して標準正規分布 $N(0,1)$ にしたがう確率変数ならば、$x=z_1^2+z_2^2+\cdots+z_n^2$ は自由度 n のカイ二乗分布にしたがう。

定理1により$z=\frac{x-\mu}{\sigma}$は標準正規分布にしたがうことが分かっているので、(45)式から$\Sigma e^2/\sigma_u^2$は自由度$(n-2)$のカイ二乗分布にしたがう。

$$\frac{e_1^2}{\sigma_u^2}+\frac{e_1^2}{\sigma_u^2}+\cdots+\frac{e_n^2}{\sigma_u^2}=\frac{\Sigma e^2}{\sigma_u^2}\sim\chi(n-2)$$

定理3：zは標準正規分布にしたがう確率変数、vは自由度ϕのカイ二乗分布にしたがう確率変数であり、さらにzとvとが互いに独立であるとき、$\frac{z\sqrt{\phi}}{\sqrt{v}}$は自由度$\phi$の$t$分布にしたがう。

傾きbについて、平均と分散はそれぞれβと$\frac{\sigma_u^2}{\Sigma x^2}$であることから、定理1により、

$$\frac{(b-\beta)}{\sigma_u/\sqrt{\Sigma x^2}}\sim N(0,1)$$

また定理3より、未知のσ_u^2を不偏推定値であるs^2に置き換えることによって正規分布からt分布へと変わる（ただしサンプル数が増えるにつれてt分布は正規分布へと近づく）。分母である$\frac{s}{\sqrt{\Sigma x^2}}$は分散の平方根であり、一般に標準誤差と呼ばれる。

$$t=z\sqrt{\phi}/\sqrt{v}=\frac{(b-\beta)}{\sigma_u/\sqrt{\Sigma x^2}}\cdot\frac{\sqrt{\Sigma e^2}/\sigma_u}{\sqrt{(n-2)}}$$
$$=\frac{(b-\beta)}{s/\sqrt{\Sigma x^2}}\sim t(n-2) \qquad (46)$$

一般的な検定手続きは(46)式によって行うことが可能である。まず傾きbの95%の信頼区間は、

$$b\pm t_{0.025}\,s/\sqrt{\Sigma x^2} \qquad (47)$$

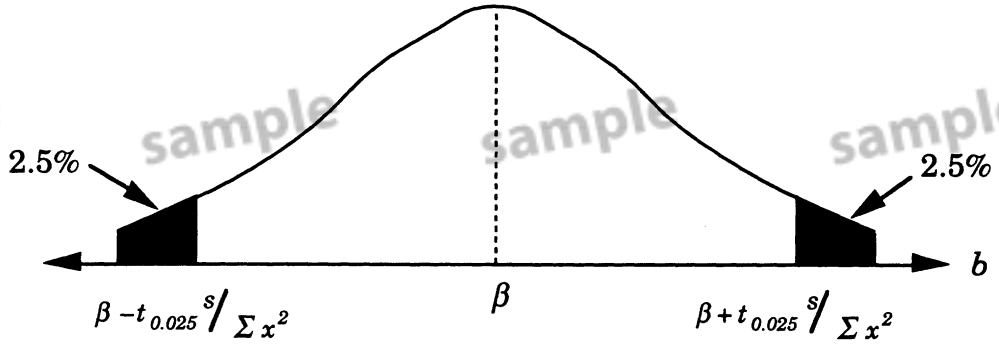
であり、図4のように描くことができる。この信頼区間とは、推定された傾きbが真の傾きであるβと95%の確率で等しい範囲を意味する。このように傾きbがβと等しいという仮説は帰無仮説(null hypothesis)と呼ばれ、次のように記述される。

H_0: $b = \beta$

逆に b が β とは異なるという仮説は対立仮説と呼ばれる。

H_1: $b \neq \beta$

（図４）信頼区間



この信頼区間は、

$$Prob\left[-t_{0.025} < \frac{(b-\beta)}{\sigma_u / \sqrt{\Sigma x^2}} < t_{0.025}\right] = 0.95$$

あるいは、

$$\left|\frac{(b-\beta)}{\sigma_u / \sqrt{\Sigma x^2}}\right| > t_{0.025}$$

と記述される。

多くの場合、帰無仮説は $\beta = 0$ であり、これを X の有意性の検定という。もし帰無仮説が有意であれば、X は Y の決定において何ら意味を持たないことになる。逆に帰無仮説が棄却された場合には、5% の水準で有意であるという。つまり Y の決定において、X の役割がゼロではないということを示すのものである。また有意性の検定には、5% や 1% といった水準が採用されるが、この数値自体には何ら根拠はなく、単に広く採用されているだけであることを認識しておくべきである。

次に切片 a についても、検定手続は傾き b と同様である。切片 a の分布は、

$$a \sim N\left(\alpha, \sigma_u^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\Sigma x^2}\right]\right)$$

であることが分かっている。したがって、

$$t = \frac{a - \alpha}{s\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\Sigma x^2}}} \sim t(n-2) \tag{48}$$

また 95% の信頼区間と帰無仮説は、それぞれ

$$a \pm t_{0.025} \cdot s\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\Sigma x^2}}$$

$$H_0 : \ a = \alpha$$

最小二乗法における分散分析

単回帰モデルに関連して、分散分析 (analysis of variance : ANOVA) が取り上げられることがある。これは (21) 式のように、全変動（総平方和）が説明された平方和と残差平方和とに分解できる特性を利用するものである。表 1 は単回帰における総平方和分解したものである。

（表 1）単回帰における分散分析

変動の原因	平方和	自由度	平均平方和
X	$ESS = \Sigma \hat{y}^2 = b^2 \Sigma x^2 = b \Sigma xy$	1	$ESS/1$
残差	$RSS = \Sigma e^2$	$n-2$	$RSS/(n-2)$
合計	$TSS = \Sigma y^2$	$n-1$	

(46) 式により

$$\frac{(b-\beta)}{\sigma_u / \sqrt{\Sigma x^2}} \sim N(0, 1)$$

したがってこれを二乗したものは、定理 2 により自由度 1 のカイ二乗分布にしたがう。

$$\frac{(b-\beta)^2}{\sigma_u^2 / \Sigma x^2} \sim \chi(1) \tag{49}$$

定理4：v_1, v_2 がそれぞれ自由度 ϕ_1, ϕ_2 のカイ二乗分布にしたがう独立した確率変数であるとき、$g = \dfrac{v_1 / \phi_1}{v_2 / \phi_2}$ は自由度 (ϕ_1, ϕ_2) の F 分布にしたがう。

定理4により、

$$F = \frac{v_1/\phi_1}{v_2/\phi_2} = \frac{(b-\beta)^2 \Sigma x^2 / 1}{\Sigma e^2/(n-2)} \sim F(1, n-2) \tag{50}$$

ここでもし帰無仮説が $\beta = 0$ であるならば、

$$F = \frac{b^2 \Sigma x^2}{\Sigma e^2/(n-2)} = \frac{b^2 \Sigma x^2}{s^2}$$

$$= \frac{\mathrm{ESS}}{\mathrm{RSS}/(n-2)} \sim F(1, n-2) \tag{51}$$

この結果を t 検定の結果と比較すれば、

$$t = \frac{(b-\beta)\sqrt{\Sigma x^2}}{s} \sim t(n-2)$$

であるため、単回帰モデルにおける F 統計量は t 統計量の二乗に等しくなる。

定理5：h が自由度 ϕ の t 分布にしたがう確率変数であるとき、h^2 は自由度 $(1, \phi)$ の F 分布にしたがう。

ここで単回帰における決定係数 R^2 と t 値、F 値との関係について簡単に述べておく。まず、$R^2 = ESS / TSS$ および $TSS = ESS + RSS$ より、

$$\frac{ESS}{RSS} = \frac{R^2}{1-R^2}$$

また、

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

であることから、

$$R^2 = \frac{F}{F + (n-2)}$$

定理５により単回帰の場合には $t^2 = F$ なので、

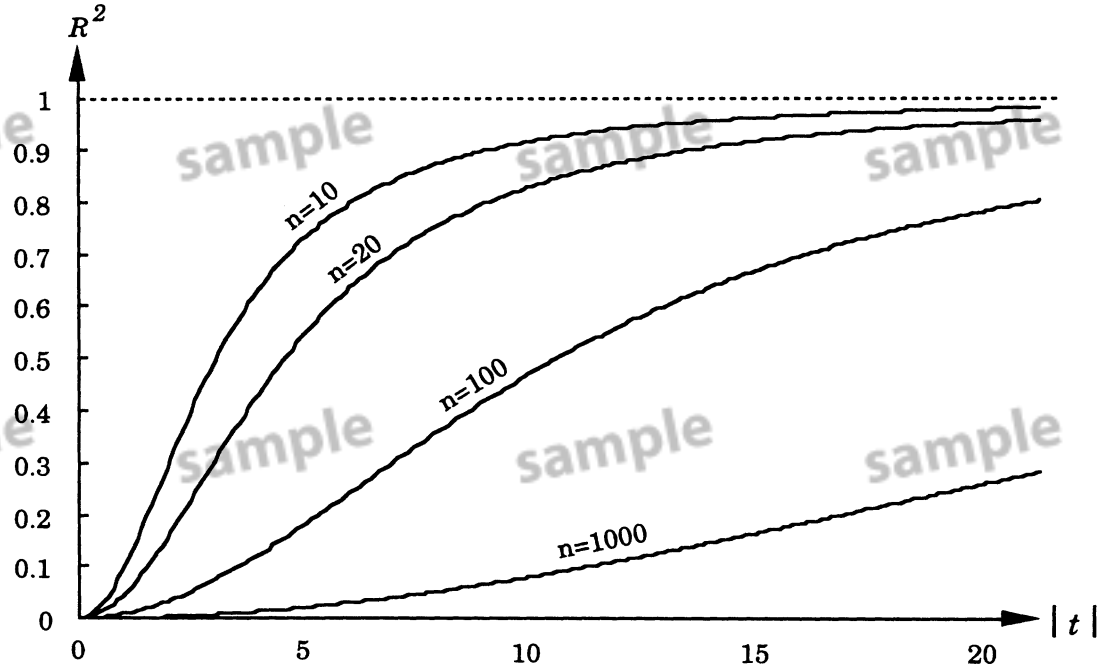
$$R^2 = \frac{t^2}{t^2 + (n-2)}$$

ここでサンプル数ごとに決定係数R^2と t 値との関係を描いたのが図５である。この図によると、サンプル数が少ないほど、小さい t 値でも高い決定係数が得られることを示している。また両者の関係式を微分すると、

$$\frac{d(R^2)}{dt} = \frac{2(n-2)t}{[t^2 + (n-2)]} \quad \text{さらに} \quad \frac{d^2(R^2)}{dt^2} = \frac{2(n-2)[(n-2) - 3t^2]}{[t^2 + (n-2)]^3} \quad \text{より、}$$

$|t| < \sqrt{\frac{n-2}{3}}$ のとき R^2は単純増加曲線、逆に $|t| > \sqrt{\frac{n-2}{3}}$ のとき R^2は単純減少曲線であることも指摘される。

（図５）決定係数R^2と t値との関係



最小二乗法における予測

これまで最小二乗法による2変数データの分析手法を考察してきた。次のステップは、新たに説明変数が観察された場合に、導き出された回帰モデルがどの程度の精度をもつのか？、またX_0を代入することによって計算された被説明変数はどのくらいの信頼区間をもつのか？、などに関連するものである。ここで任意の値(X_0)についての回帰式の予測を$\left(\hat{Y}_0\right)$とすると、

$$\hat{Y}_0 = a + b\,X_0 \tag{52}$$

また、

$$Y_0 = \alpha + \beta\,X_0 + u_0 \tag{53}$$

によって予測誤差(e_0)は、

$$\begin{aligned} e_0 &= Y_0 - \hat{Y}_0 \\ &= u_0 - (a-\alpha) - (b-\beta)\,X_0 \end{aligned} \tag{54}$$

またaとbはそれぞれαとβの不偏推定値であるため、$E(a-\alpha)=0$、$E(b-\beta)=0$である。さらに$E(u_0)=0$により、予測誤差の期待値は$E(u_0)=0$である。これは(45)式による予測値が不偏であることを意味する。また予測誤差の分散については、

$$\begin{aligned} Var(e_0) &= E(e_0^2) \\ &= E\left[u_0^2 + (a-\alpha)^2 + (b-\beta)^2 X_0^2 - 2(a-\alpha)u_0 \right. \\ &\qquad \left. - 2(b-\beta)X_0 u_0 - 2(a-\alpha)(b-\beta)X_0\right] \\ &= Var(u_0^2) + Var(a) + X_0^2\,Var(b) - 2\,X_0\,cov(a,b) \\ &= \sigma_u^2 + \sigma_u^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\Sigma x^2}\right] + X_0^2\frac{\sigma_u^2}{\Sigma x^2} - 2\,X_0\,\sigma_u^2\frac{\bar{X}}{\Sigma x^2} \\ &= \sigma_u^2\left[1 + \frac{1}{n} + \frac{(X_0-\bar{X})^2}{\Sigma x^2}\right] \end{aligned} \tag{55}$$

つまりX_0が$\bar{X}$に等しいとき、誤差の分散$\sigma_u^2\left[1+\frac{1}{n}\right]$は最小になる。また(54)式より$e_0$は線形関数であり、さらに正規分布にしたがうので、

$$z = \frac{e_0}{\sigma} = \frac{e_0}{\sigma_u \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma x^2}}} \sim N(0, 1) \tag{56}$$

ここで未知である σ_u^2 を推定値 $s^2 = \Sigma e^2 / (n-2)$ に置き換えることによって、e_0 は自由度が $n-2$ の t 分布にしたがうことになる。

$$t = \frac{Y_0 - \hat{Y}_0}{s \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma x^2}}} \sim t(n-2) \tag{57}$$

以上により、未知の数値である Y_0 の 95% 水準の信頼区間は、

$$(a + b X_0) \pm t_{0.025} \cdot s \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma x^2}} \tag{58}$$

次に、もし Y_0 自体の数値よりも Y_0 の 平均について興味があるのであれば、(53) 式の期待値をとればよい。

$$E(Y_0) = \alpha + \beta X_0 \tag{59}$$

したがって、

$$\begin{aligned} e_0 &= E(Y_0) - \hat{Y}_0 \\ &= -(a - \alpha) - (b - \beta) X_0 \end{aligned} \tag{60}$$

この分散は、

$$\begin{aligned} Var(e_0) &= E\left[(a - \alpha)^2 + X_0^2 (b - \beta)^2 - 2 (a - \alpha)(b - \beta) X_0\right] \\ &= Var(a) + X_0^2 Var(b) - 2 X_0 cov(a, b) \\ &= \sigma_u^2 \left[\frac{1}{n} + \frac{(\bar{X} - X_0)}{\Sigma x^2}\right] \end{aligned} \tag{61}$$

つまり Y_0 の平均である $E(Y_0)$ の 95% 水準の信頼区間は、

$$(a + b X_0) \pm t_{0.025} \cdot s \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma x^2}} \tag{62}$$

によって求めることができる。

例題

これまでの単回帰の仮定や手法を数学的に考察してきたが、例題を用いて実際に手計算することによって最小二乗法の仕組みを理解することが望ましい。表2は、ある機械の稼働時間 (X) と固定費 (Y) とを 30 ヶ月 (n) にわたって測定したものである。ここでの興味は、固定費を決定する上で機械の稼働時間がどのような影響を及ぼすのか？、またそれは確率的に有意なのか？、などに関連するものである。

（表2）機械の稼働時間と固定費

月数 (i)	固定費 (X)	機械の稼働時間 (Y)
1	76.67	1.77
2	73.68	1.82
3	80.14	1.63
4	61.99	1.01
5	72.69	1.38
6	87.68	1.96
7	78.45	1.56
8	70.63	1.46
9	63.42	1.55
10	56.06	1.12
11	67.45	1.38
12	72.10	1.32
13	68.53	1.26
14	69.08	1.34
15	85.55	1.80
16	58.20	1.02
17	61.63	1.51
18	80.69	1.79
19	58.26	1.15
20	55.34	1.16
21	85.11	1.85
22	76.49	1.83
23	67.78	1.14
24	56.40	1.14
25	66.62	1.33
26	63.49	1.36
27	89.42	1.88
28	67.52	1.17
29	73.68	1.64
30	66.13	1.38

(Kaplan,R.S & A.A.Atkinson "Advanced Management Accounting", Prentice-Hall, N.Y: 1989; pp.99.を修正)

$$\Sigma X = 43.71 \qquad \Sigma Y = 2,110.88 \qquad \Sigma XY = 3,144.157$$
$$\Sigma X^2 = 66.039$$

これらを(8) 式と (9) 式に代入すれば、

$$2,110.88 = 30a + b\,(43.71)$$
$$3,144.157 = a\,(43.71) + b\,(66.039)$$

この方程式を解くと、$a = 27.891$、$b = 29.150$ が求められる。この結果は、機械を稼働させなくとも固定費が 27.891 発生し、また稼働時間が 1 時間延びるにつき固定費が 29.150 発生するというものである。別解としては、(11) 式に

$$\Sigma x^2 = 2.354 \qquad \Sigma xy = 68.605$$

を代入すれば切片 b が導かれる。

$$b = \Sigma xy \Big/ \Sigma x^2 = 68.605 \Big/ 2.354 = 29.144 \quad \text{(A)}$$

さらに切片 a は (10) 式により求めることができる。

$$a = \bar{Y} - b\,\bar{X} = 70.363 - (29.144)(1.457) = 27.900 \quad \text{(B)}$$

ここで、$\bar{X} = \Sigma X \Big/ n = 1.457$ $\qquad$ $\bar{Y} = \Sigma Y \Big/ n = 70.363$

次に、この回帰モデルの総平方和 (TSS) を説明された平方和 (ESS) と残差平方和 (RSS) とに分解すると、

$$\text{TSS} = \Sigma y^2 = 2{,}708.273$$

$$\text{ESS} = \Sigma \hat{y}^2 = (\Sigma xy)^2 \Big/ \Sigma x^2 = (68.605)^2 \Big/ 2.354 = 1{,}999.425 \quad \text{(C)}$$

$$\text{RSS} = \text{TSS} - \text{ESS} = 2{,}708.273 - 1{,}999.425 = 708.849 \quad \text{(D)}$$

残差平方和 (RSS) を検算する方法としては、サンプルごとに被説明変数の推定値を測定し、その平方和 $\left[\Sigma (Y - a - b\,X)^2\right]$ を計測すればよい。このように測定された 3 つの平方和により、決定係数と相関係数とが導き出される。まず (24) 式によって決定係数は、

$$r^2_{xy} = \text{ESS} \Big/ \text{TSS} = 1{,}999.425 \Big/ 2{,}708.273 = 0.738 \quad \text{(E)}$$

これは、機械の稼働時間によって固定費の全変動の 73.8% が説明されることを意味する。また機械の稼働時間と固定費との直線関係の尺度である相関係数は、

$$r_{xy} = \sqrt{0.738} = 0.859 \quad \text{(F)}$$

であり、両変数の関係が究めて直線に近いことを示している。

以上の分析によって、切片 a と傾き b の平均値を測定したが、次の段階はこれらの分散についてである。まず (38) 式により切片 a の分散は、

$$Var(a) = \sigma_u^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\Sigma x^2} \right) = \frac{\Sigma e^2}{n-2} \left(\frac{1}{n} + \frac{\bar{X}^2}{\Sigma x^2} \right) = \frac{708.849}{28} \left[\frac{1}{30} + \frac{(1.457)^2}{2.354} \right] = 23.674$$

また (39) 式により傾き b の分散は、

$$Var(b) = \frac{\sigma_u^2}{\Sigma x^2} = \frac{\Sigma e^2}{n-2} \cdot \frac{1}{\Sigma x^2} = \frac{708.849}{28} \cdot \frac{1}{2.354} = 10.754$$

つまり切片 a と傾き b は、それぞれ次のような平均値と分散をもつ正規分布にしたがう。

$$a = N(27.900,\ 23.674) \qquad b = N(29.144,\ 10.754)$$

またこれらの 95% の信頼区間は (34) 式と (35) 式によって求めることができる。付属資料１のスチューデントの t 分布表によると、自由度が２８のときの $t_{0.025}$ は 2.048 である。したがってこの t 値を (34) 式と (35) 式とに代入すると、

$$b \pm t_{0.025} \frac{s}{\sqrt{\Sigma x^2}} = b \pm t_{0.025} \sqrt{Var(b)} = 29.144 \pm 2.048 \sqrt{10.754}$$
$$= [22.428,\ 35.860]$$

$$a \pm t_{0.025} \frac{s}{\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\Sigma x^2}}} = a \pm t_{0.025} \sqrt{Var(a)} = 27.900 \pm 2.048 \sqrt{23.674}$$
$$= [17.935,\ 37.864]$$

これは95%の確率で、真の傾きがこの範囲に存在するというものである。しかし信頼期間は広く、意思決定には何ら有用ではない。

次に切片と傾きの t 値を計算すると、それぞれ

$$t = \frac{b}{\frac{s}{\sqrt{\Sigma x^2}}} = \frac{b}{\sqrt{Var(b)}} = \frac{29.144}{\sqrt{10.754}} = 8.887 \qquad \text{(G)}$$

$$t = \frac{a}{\frac{s}{\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\Sigma x^2}}}} = \frac{a}{\sqrt{Var(a)}} = \frac{27.900}{\sqrt{23.674}} = 5.734 \qquad \text{(H)}$$

この結果から理解されるように、切片 a と傾き b はともに $t_{0.025}$ である 2.048 、さらに $t_{0.005}$ である 2.763 より大きい。したがって、回帰モデルによって導き出された切片と傾きがゼロであるという帰無仮説は、99% の水準で棄却されることになる。

つまり固定費の決定に機械の稼働率はプラスの影響を与えていることが否定できないことが検証されたのである。次に、分散分析によるF検定によってこの結果を確かめてみる。(51) 式により、傾き b は自由度 (1, 28) の F 値にしたがうことから、

$$F = \frac{\mathrm{ESS}}{\mathrm{RSS}/_{n-2}} = \frac{1{,}999.425}{708.849/_{30-2}} = 78.979 \sim F(1,\ 28) \qquad \text{(I)}$$

この F 値の平方根はt 値と同じ 8.887 である。

次に、この回帰モデルを用いて予測を行う場合の留意点について述べることにする。もし経営者や固定費の管理責任者が、ある特定の稼働時間 (X_0) がどの程度の固定費 (Y_0) を発生させるかについて予測を行うものと仮定しよう。この場合の予測とは、区間予測と点（平均）予測の 2 つである。ここで想定される稼働時間 (X_0) を 1.50 と 2.00 時間とする。前者は標本内予測であるが、後者の水準は観察データを越えたものであり、これを標本外予測と呼ぶ。まず稼働時間 (X_0) が 1.50 の場合についての予測を考えてみると、予測誤差の分散は (55) 式によって求めることができる。

$$Var(e_0) = \sigma_u^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma x^2}\right] = \frac{708.849}{28}\left[1 + \frac{1}{30} + \frac{(1.50 - 1.457)^2}{2.354}\right] = 26.180$$

この分散の数値を (58) 式に代入すると、次のような信頼区間が得られる。

$$(a + b\,X_0) \pm t_{0.025}\sqrt{Var(e_0)} = [27.900 + (29.144)(1.50)] \pm 2.048\sqrt{26.180}$$
$$= [61.137,\ 82.095]$$

また稼働時間 (X_0) が 2.00 の場合については、[75.709, 96.667] の範囲が 95% の信頼区間となる。このように稼働時間の平均に対して、より乖離が大きい 2.00 時間の予測の方が信頼区間の範囲が広くなる。しかし予測者、もしくは分析者の関心は、信頼区間の計測よりも予測点や予測平均に向けられることが多い。この平均の予測についての 95% の信頼区間は、(62) 式により

$$(a + b\,X_0) \pm t_{0.025}\,s\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma x^2}} = [69.711,\ 73.521]$$

および、[*82.080*, *90.296*] である。この場合にも、平均より乖離するほどその範囲は広くなり、意思決定に際して不確定要素が大きくなることを意味している。

表3は、この例題の統計ソフトによる回帰分析の結果である。使用したパッケージはSYSTATバージョン5.2.1であり、この例題における結果を記号(A)、(B)、(C)…(I)で示している。

（表３）SYSTAT バージョン 5.2.1 による分析結果

DEP VAR: Overhead N: 30
MULTIPLE R: 0.859 [F]
SQUARED MULTIPLE R: 0.738 [E]
ADJUSTED SQUARED MULTIPLE R: 0.729
STANDARD ERROR OF ESTIMATE: 5.032

VARIABLE	COEFFICIENT	STD ERROR	STD COEF	T	P(2 TAIL)
CONSTANT	27.900 [B]	4.866	0.000	5.734 [H]	.38E-05
Machine-Hours	29.144 [A]	3.279	0.859	8.887 [G]	.12E-08

ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	1999.413 [C]	1	1999.413	78.977 [I]	.121767E-08
RESIDUAL	708.859	28	25.316		