# 多変量解析　目次

１．回帰分析

何名かの体重と身長の値が分かっているとき、体重の値は分かっているが、身長が不明の人がいるとする。このようなとき、すでに得ているデータから身長と体重の関係を調べ、その相関を求め、身長不明の人の身長を予測する。この様な分析方法を回帰分析という。

求めるものは身長であり、これを目的変量と呼ぶ。身長の値を予測するのは、体重からであるので、この体重のことを説明変量と呼ぶ。説明変量が１つの時を単回帰分析といい、説明変量が２つ以上の時を多重回帰分析という。

回帰分析では、説明変量は量的データであり、また目的変量も量的データである。

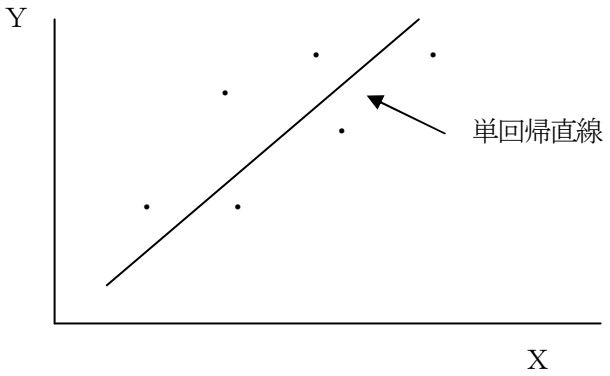
なお、回帰式で予測をするときには、説明変量の範囲内で予測することが望ましい。説明変量の範囲を大きく越えたところで予測すると誤差が大きくなり実用に適さなくなる。

１．１　単回帰分析

正規母集団から抽出して得られた標本データｘ・ｙが下表のようにあり、ｘ・ｙ間にある関係があるものとする。

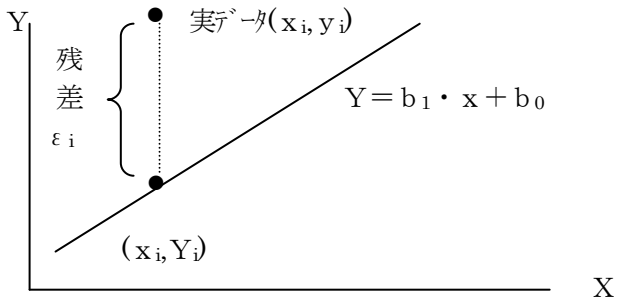
| 標　本 | 説明変量ｘ | 目的変量ｙ |
|---|---|---|
| 1 | $x_1$ | $y_1$ |
| 2 | $x_2$ | $y_2$ |
| … | … | … |
| n | $x_n$ | $y_n$ |

以上の標本データをＸＹグラフで描くいて、下のようになったとする




標本データｘ・ｙの間には右上がりの関係がありそうなので、ｘとｙの関係を表す適当な直線を考える。 目的変量ｙと説明変量ｘとの間に相関があるとき、

$Y = b_1 \cdot x + b_0$　なる直線を１本考え、実データとこの直線上の値との差を $\varepsilon$ とする。




$Y = b_1 \cdot x + b_0$　なる直線は全ての標本データについて、その残差が最小になるようにひく必

要がある。この直線から各標本データとのズレ具合いを計るために、各残差の平方和をとり、この平方和を最小にするようにする。このような方法を最小２乗法という。

標本データは、直線$Y=b_1\cdot x+b_0$ から残差（$\varepsilon$）分ずれているので、標本データは

$y=b_1\cdot x+b_0+\varepsilon$ と表す。

このことから線形回帰モデルを

$y_i=\beta_1\cdot x_i+\beta_0+\varepsilon_i$　（i=1,2…n）とすると

残差$\varepsilon$について、

①$\varepsilon_i$と$\varepsilon_j$はお互いに独立であり、正規分布　N（0，$\sigma^2$）に従う。
②$\varepsilon_i$の平均値（期待値）は0である。
③$\varepsilon_i$の分散は一定である。

このような仮定下で単回帰式を $Y=b_1\cdot x+b_0$ とする。

いま、残差$\varepsilon$に注目すると　$\varepsilon_i=y_i-Y_i$　$\varepsilon_i=y_i-b_1\cdot x_i-b_0$　である。

この残差を全ての標本データについて合計し、その合計値を最小にするような$b_0\cdot b_1$　を求め、この単回帰式を得る。

$\Sigma\varepsilon_i^2=\Sigma(y_i-b_1\cdot x_i-b_0)^2$ であるから

$f=\Sigma(y_i-b_1\cdot x_i-b_0)^2$とすると

この式を$b^0$，$b_1$で偏微分して、0とおくことにより、正規方程式を得て、式 f を最小 にする$b_0\cdot b_1$を得ることができる。

$$\begin{cases}\dfrac{\partial f}{\partial b_1}=-2\Sigma x_i\cdot(y_i-b_1\cdot x_i-b_0)=0\\[2ex]\dfrac{\partial f}{\partial b_0}=-2\Sigma(y_i-b_1\cdot x_i-b_0)=0\end{cases}$$

これから

$$\begin{cases}b_1=\dfrac{n\cdot\Sigma x_i\cdot y_i-\Sigma x_i\cdot\Sigma y_i}{n\cdot\Sigma x_i^2-(\Sigma x_i)^2}\\[2ex]b_0=\dfrac{\Sigma x_i^2\cdot\Sigma y_i\text{ -}-\Sigma x_i\cdot y_i\cdot\Sigma x_i}{n\cdot\Sigma x_i^2-(\Sigma x_i)^2}\end{cases}$$

また

$$b_1=\frac{n\cdot\Sigma x_i\cdot y_i-\Sigma x_i\cdot\Sigma y_i}{n\cdot\Sigma x_i^2-(\Sigma x_i)^2}=\frac{\Sigma x_i\cdot y_i-\dfrac{\Sigma x_i\cdot y_i}{n}}{\Sigma x_i^2-\dfrac{(\Sigma x_i)^2}{n}}=\frac{\Sigma(x_i-\overline{x})\cdot(yi-\overline{y})}{\Sigma(x_i-\overline{x})^2}=\frac{Sxy}{Sxx}$$

ただし、xi・yiの偏差平方和・積和をそれぞれ Sxx・Syy・Sxy とすると

$Sxx=\Sigma(x_i-\overline{x})^2$　　$Syy=\Sigma(y_i-\overline{y})^2$

$Sxy=\Sigma(x_i-\overline{x})\cdot(y_i-\overline{y})$ である。

以上から、単回帰式は

$Y-\overline{y}=b_1(x-\overline{x})$　→　$Y=\dfrac{Sxy}{Sxx}(x-\overline{x})+\overline{y}$　と表される。

またｘとｙの相関係数を $R_{xy}$ とすると

$$R_{xy}=\frac{Sxy}{\sqrt{Sxx}\sqrt{Syy}}$$ であるから、

相関係数を使用して単回帰式を表すと

$$Y=R\cdot\frac{\sqrt{Syy}}{\sqrt{Sxx}}(x-\overline{x})+\overline{y}$$ と書くことができる。

１．２　重回帰分析

それでは次に説明変量が $x_1$・$x_2$ の２変量になったときの回帰式を求める。

| 標　本 | 説明変量 $x_1$ | 説明変量 $x_2$ | 目的変量ｙ |
|---|---|---|---|
| 1 | $x_{11}$ | $x_{21}$ | $y_1$ |
| 2 | $x_{12}$ | $x_{22}$ | $y_2$ |
| … | … | … | … |
| n | $x_{1n}$ | $x_{2n}$ | $y_n$ |

説明変量が２変量あるので、単純に説明２変量（$x_1$ と $x_2$）の平均値をとって、その値と目的変量（ｙ）との相関を求めても、平均値をとる段階で失う情報量が大きいので正しい回帰式を得ることができない。

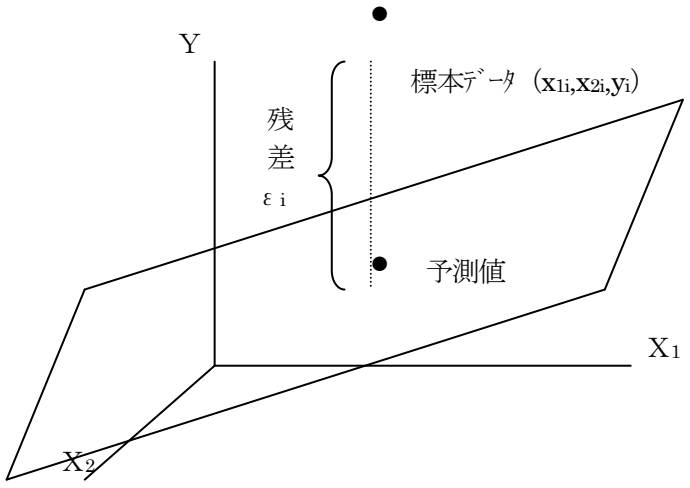
このように説明変量が２つ以上ある時の回帰分析を重回帰分析という。

1.2.1　重回帰式を求める。

２説明変量が次のようになっているときの重回帰直線を求める。

| 標　本 | 説明変量 $x_1$ | 説明変量 $x_2$ | 目的変量ｙ |
|---|---|---|---|
| 1 | $x_{11}$ | $x_{21}$ | $y_1$ |
| 2 | $x_{12}$ | $x_{22}$ | $y_2$ |
| … | … | … | … |
| n | $x_{1n}$ | $x_{2n}$ | $y_n$ |
| 平　均 | $\overline{x_1}$ | $\overline{x_2}$ | $\overline{y}$ |

この関係を図で表すと



説明変量（$x_1$，$x_2$）と目的変量（y）との間に相関関係があるとき

$$Y = b_1 \cdot x_1 + b_2 \cdot x_2 + b_0$$

なる平面を考え、実際の標本データからこのこの平面上への残差を$\varepsilon$とすると

説明変量が２つある時の重回帰式は

$$y_i = b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + b_0 + \varepsilon_i \quad \text{と表される。}$$

残差$\varepsilon$に注目すると　$\varepsilon_i = y_i - Y_i$

$$\varepsilon_i = y_i - (b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + b_0) \quad \text{であるから}$$

この残差平方和を求め、残差平方和が最小にするような$b_0 \cdot b_1 \cdot b_2$を求めると、重回帰式を得ることができる。

一般に説明変量がｐ個ある時の線形重回帰モデルは

$$y_i = \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \cdots + \beta_p \cdot x_{pi} + \beta_0 + \varepsilon_i \quad (i=1,2 \cdots n)$$

と表される。この時単回帰分析と同様に

残差$\varepsilon$について、

①$\varepsilon_i$と$\varepsilon_j$はお互いに独立であり、正規分布　N（０，$\sigma^2$）に従う。

②$\varepsilon_i$の平均値（期待値）は０である。

③$\varepsilon_i$の分散は一定である。

との仮定下で重回帰予測式を

$Y_i = b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + \cdots + b_p \cdot x_{pi} + b_0$ とする。

$b_1 \cdot b_2 \cdots b_p$を偏回帰係数といい、　$\beta_1 \cdot \beta_2 \cdots \beta_p$ を母偏回帰係数という。

［残差平方和$\Sigma (\varepsilon_i)^2$ を最小にするような$b_0 \cdot b_1 \cdot b_2$ を求める。］

$\Sigma (\varepsilon_i)^2 = \Sigma \{y_i - (b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + b_0)\}^2$ を最小にする$b_0 \cdot b_1 \cdot b_2$ を 求める。

$f = \Sigma (y_i - b_1 \cdot x_{1i} - b_2 \cdot x_{2i} - b_0)^2$ とし、この式を$b_0 \cdot b_1 \cdot b_2$で偏微分する。

$$\begin{cases} \dfrac{\partial f}{\partial b_1} = -2 \ \Sigma x_{1i} \cdot (y_i - b_1 \cdot x_{1i} - b_2 \cdot x_{2i} - b_0) = 0 \\ \dfrac{\partial f}{\partial b_1} = -2 \ \Sigma x_{2i} \cdot (y_i - b_1 \cdot x_{1i} - b_2 \cdot x_{2i} - b_0) = 0 \\ \dfrac{\partial f}{\partial b_1} = -2 \ \Sigma (y_i - b_1 \cdot x_{1i} - b_2 \cdot x_{2i} - b_0) = 0 \end{cases}$$

これより

$$\begin{cases} \Sigma x_{1i} \cdot (y_i - b_1 \cdot x_{1i} - b_2 \cdot x_{2i} - b_0) = 0 & \cdots① \\ \Sigma x_{2i} \cdot (y_i - b_1 \cdot x_{1i} - b_2 \cdot x_{2i} - b_0) = 0 & \cdots② \\ \Sigma (y_i - b_1 \cdot x_{1i} - b_2 \cdot x_{2i} - b_0) = 0 & \cdots③ \end{cases}$$

上の式を正規方程式という

③から　$\Sigma y_i - b_1 \cdot \Sigma x_{1i} - b_2 \cdot \Sigma x_{2i} - \Sigma b_0 = 0$

$\Sigma b_0 = n \cdot b_0$であるから

$$b_0 = \frac{\Sigma y_i - b_1 \cdot \Sigma x_{1i} - b_2 \cdot \Sigma x_{2i}}{n} = \overline{y} - b_1 \cdot \overline{x_1} - b_2 \cdot \overline{x_2}$$

これを①②に代入して

$$b_1 \cdot (\Sigma x_{1i}^2 - n \cdot \overline{x_1}^2) + b_2 \cdot (\Sigma x_{1i} \cdot x_{2i} - n \cdot \overline{x_1} \cdot \overline{x_2}) = \Sigma x_{1i} \cdot y_i - n \cdot \overline{x_1} \cdot \overline{y}$$

$$b_1 \cdot (\Sigma x_{1i} x_{2i} - n \cdot \overline{x_1} \cdot \overline{x_2}) + b_2 \cdot (\Sigma x_{2i}^2 - n \cdot \overline{x_2}) = \Sigma x_{2i} \cdot y_i - n \cdot \overline{x_2} \cdot \overline{y}$$

これより$b_0 \cdot b_1 \cdot b_2$ を求めると、重回帰式の係数を得ることができる。

1.2.2　偏差平方和・積和から重回帰式を求める

⑴説明変量が２個の時

説明変量$x_1$・$x_2$ の偏差平方和それぞれ$S_{11}$・$S_{22}$、偏差積和を$S_{12}$とすると

$$\begin{cases} S_{11} = \Sigma (x_{1i} - \overline{x_1})^2 = \Sigma x_{1i}^2 - n \cdot \overline{x_1} \\ S_{22} = \Sigma (x_{2i} - \overline{x_2})^2 = \Sigma x_{2i}^2 - n \cdot \overline{x_2} \\ S_{12} = \Sigma (x_{1i} - \overline{x_1}) \cdot (x_{2i} - \overline{x_2}) = \Sigma x_{1i} \cdot x_{2i} - n \cdot \overline{x_1} \cdot \overline{x_2} \end{cases}$$

また目的変量ｙと説明変量$x_1$との偏差積和を$S_{y1}$とすると

$$S_{y1} = \Sigma (x_{1i} - \overline{x_1}) \cdot (y_i - \overline{y}) = \Sigma x_{1i} \cdot y_i - n \cdot \overline{x_1} \cdot \overline{y}$$

目的変量ｙと説明変量$x_2$との偏差積和を$S_{y2}$とすると

$$S_{y2} = \Sigma (x_{2i} - \overline{x_2}) \cdot (y_i - \overline{y}) = \Sigma x_{2i} \cdot yi - n \cdot \overline{x_2} \cdot \overline{y}$$

以上から前の式は

$$\begin{cases} S_{11} \cdot b_1 + S_{12} \cdot b_2 = S_{y1} \\ S_{12} \cdot b_1 + S_{22} \cdot b_2 = S_{y2} \end{cases}$$

となるので、これから係数$b_0$・$b_1$・$b_2$を求める。

また$b_0 = \overline{y} - (b_1 \cdot \overline{x_1} + b_2 \cdot \overline{x_2})$である。

⑵説明変量がｐ個ある時

平方和・積和行列を

$$\mathbf{S} = \begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{pmatrix}$$

とする。

求める重回帰式を、$Y_i = b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + \cdots + b_p \cdot x_{pi} + b_0$　とする、この回帰式の係数$b_0$・$b_1 \cdots b_p$ は、下の連立方程式の解として与えられる。

$$\begin{cases} S_{11} \cdot b_1 + S_{12} . b_2 + \cdots + S_{1p} \cdot b_p = S_{y1} \\ S_{21} \cdot b_1 + S_{22} . b_2 + \cdots + S_{2p} \cdot b_p = S_{y2} \\ \cdots \\ S_{p1} \cdot b_1 + S_{p2} \cdot b_2 + \cdots + S_{pp} \cdot b_p = S_{yp} \end{cases}$$

係数$b_0$は、　$b_0 = \overline{y} - (b_1 \cdot \overline{x_1} + b_2 \cdot \overline{x_2} + \cdots + b_p \cdot \overline{x_P})$ である。

また行列では

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ \cdots \\ y_{1n} \end{pmatrix} \quad \mathbf{x} = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ & \cdots & & & \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \cdots \\ b_n \end{pmatrix}$$

とすると

係数ｂは、　$\mathbf{b} = (\mathbf{x}' \cdot \mathbf{x})^{-1} \cdot \mathbf{x}' \cdot \mathbf{y}$　で求めることができる。

## １．３　標準偏回帰係数

説明変量がどれくらい目的変量に影響を与えているか（寄与しているか）を見るには、求めた重回帰式の偏回帰係数を見ればよい。通常、偏回帰係数が大きいほど目的変量に与える影響が大きいので多く寄与しているといえる。しかし、説明変量間で単位が異なるときには、単位の影響を受けるので､単純に偏回帰係数の大小比較して決めることはできない｡単位の影響を除くには、標本データを標準化する。データを標準化することにより、平均＝０・分散＝１となり単位の影響を受けなくなるので､標準化したデータから偏回帰 係数を求めるようにする｡このように標準化したデータから得られた偏回帰係数を、標準 偏回帰係数という。

標準偏回帰係数の大きいほど、目的変量に与える影響が大きく、寄与の大きい変量であるといえる。

通常説明変量が２つの時の重回帰式は

$Y-\overline{y} = b_1(x_1-\overline{x_1}) + b_2(x_2-\overline{x_2})$ と書ける。

いま、目的変量の標準偏差を$\sqrt{S_{yy}}$ 、説明変量$x_1$の標準偏差を$\sqrt{S_{11}}$ 、説明変量$x_2$の標準偏差を$\sqrt{S_{22}}$ とすると

データの標準化は

$$Y \rightarrow \frac{Y-\overline{y}}{\sqrt{S_{yy}}} \qquad x_1 \rightarrow \frac{x_1-\overline{x_1}}{\sqrt{S_{11}}} \qquad x_2 \rightarrow \frac{x_2-\overline{x_2}}{\sqrt{S_{22}}}$$ を行うことである。

$$\frac{Y-\overline{y}}{\sqrt{S_{yy}}} = b_1\cdot\frac{x_1-\overline{x_1}}{\sqrt{S_{yy}}} + b_2\cdot\frac{x_2-\overline{x_2}}{\sqrt{S_{yy}}} = b_1\cdot\frac{\sqrt{S_{11}}}{\sqrt{S_{yy}}}\cdot\frac{x_1-\overline{x_1}}{\sqrt{S_{11}}} + b_2\cdot\frac{\sqrt{S_{22}}}{\sqrt{S_{yy}}}\cdot\frac{x_2-\overline{x_2}}{\sqrt{S_{22}}}$$

データを標準化して得られる重回帰式の係数$b_1$'・$b_2$'は

$$b_1' = b_1\cdot\frac{\sqrt{S_{11}}}{\sqrt{S_{yy}}} \qquad b_2' = b_2\cdot\frac{\sqrt{S_{22}}}{\sqrt{S_{yy}}}$$ と表すことができる。

## １．４　相関係数と決定係数

### 1.4.1　単回帰式における相関係数と決定係数

| 標　本 | 説明変量x | 目的変量y |
|---|---|---|
| 1 | $x_1$ | $y_1$ |
| 2 | $x_2$ | $y_2$ |
| … | … | … |
| n | $x_n$ | $y_n$ |
| 平　均 | $\overline{x}$ | $\overline{y}$ |

説明変量ｘの変化に従って目的変量ｙが変化する（相関関係にある）ときｘとｙの間の相関係数をＲとすると

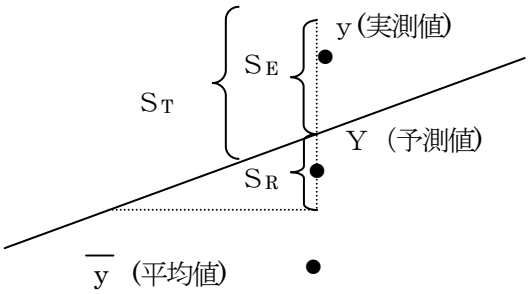
$$R=\frac{\Sigma(x_i-\overline{x})\cdot(y_i-\overline{y})}{\sqrt{\Sigma(x_i-\overline{x})^2}\cdot\sqrt{\Sigma(y_i-\overline{y})^2}}=\frac{Sxy}{\sqrt{Sxx}\sqrt{Syy}}$$

相関係数Ｒは　－１≦Ｒ≦１　の値をとり

- Ｒ＞０　…　正の相関がある
- Ｒ＝１　…　無相関
- Ｒ＜０　…　負の相関がある

いま説明変量ｘと実測値ｙとの関係がｒである時、これから求めた単回帰式を

Ｙ＝ｂ１・ｘ＋ｂ０ とすると




実測値ｙは、単回帰直線の付近にばらついて散在している。このばらつきの小さいほど単回帰式のあてはまりがよい（精度が高い）直線といえる。また説明変量ｘの目的変量に与える影響が大きいといえる。つまり決定力が大きいといえる。

分散状況を見ると、全分散（$S_T$）は、実測値$y_i$が平均値ｙからどれ位分散しているかであるので、$\Sigma(y_i-\overline{y})^2$。回帰で説明可能な部分の分散（$S_R$）、つまり予測値が平均値からどれ位分散しているかは、$\Sigma(Y_i-\overline{y})^2$。回帰で説明できない残差部分の分散（$S_E$）つまり実測値が予測値からどれ位分散しているかは、$\Sigma(y_i-Y_i)^2$である。

これらの変動の間には、

$S_T=S_R+S_E$ つまり $\Sigma(y_i-\overline{y})^2=\Sigma(Y_i-\overline{y})^2+\Sigma(y_i-Y_i)^2$ なる関係がある。

この両辺を$\Sigma(y_i-\overline{y})^2$で割ると

$$1=\frac{\Sigma(Y_i-\overline{y})^2}{\Sigma(y_i-\overline{y})^2}+\frac{\Sigma(y_i-Y_i)^2}{\Sigma(y_i-\overline{y})^2}\qquad 1-\frac{\Sigma(y_i-Y_i)^2}{\Sigma(y_i-\overline{y})^2}=\frac{\Sigma(Y_i-\overline{y})^2}{\Sigma(y_i-\overline{y})^2}=\frac{S_R}{S_T}=R^2$$

この$R^2$のことを決定係数という。この決定係数は　０≦$R^2$≦１の値をとる。
また、この決定係数$R^2$は相関係数Ｒの２乗に等しい。

1.4.2　重回帰式における相関係数と決定係数

（1）重相関係数と決定係数

| 標本No | 説明変量 $x_1$ | $x_2$ | … | $x_p$ | 実測値 $y$ | 予測値 $Y$ |
|---|---|---|---|---|---|---|
| 1 | $x_{11}$ | $x_{21}$ | … | $x_{p1}$ | $y_1$ | $Y_1$ |
| 2 | $x_{12}$ | $x_{22}$ | … | $x_{p2}$ | $y_2$ | $Y_2$ |
| … | | … | | | … | … |
| n | $x_{1n}$ | $x_{2n}$ | … | $x_{pn}$ | $y_n$ | $Y_n$ |
| 平　均 | $\overline{x_1}$ | $\overline{x_2}$ | … | $\overline{x_P}$ | $\overline{y}$ | $\overline{Y}$ |

重相関係数Rは、実測値データｙと重回帰式から求めた予測値データＹとの相関係数である。

$$R=\frac{\Sigma(y_i-\overline{y})\cdot(Y_i-\overline{y})}{\sqrt{\Sigma(y_i-\overline{y})^2}\cdot\sqrt{\Sigma(Y_i-\overline{y})^2}}=\sqrt{\frac{\Sigma(Yi-\overline{y})^2}{\Sigma(yi-\overline{y})^2}}$$

また単回帰のときと同様に、相関係数の２乗を決定係数と呼び、やはり０≦$R^2$≦１の値をとる。$R^2$ が１に近いほど重回帰式の精度が高いといえる。

$$R^2=1-\frac{\Sigma(y_i-Y_i)^2}{\Sigma(y_i-\overline{y})^2}=\frac{\Sigma(Y_i-\overline{y})^2}{\Sigma(y_i-\overline{y})^2}$$

［重相関係数の検定］

標本から得られた重相関係数について、その母重相関係数（$\rho$）が無相関かどうかの検定を行う。標本から得られた重相関係数をＲとする時、その母相関係数（$\rho$）について$\rho=0$の仮説につき、検定統計量をＦとすると

$$F=\frac{R^2/p}{(1-R^2)/(n-p-1)}$$　ただし　ｐ：説明変量の個数　ｎ：標本数　Ｒ：重相関係数

は、自由度ｐ，ｎ－ｐ－１のＦ分布に従うことを利用して検定を行う。

検定をおこなう

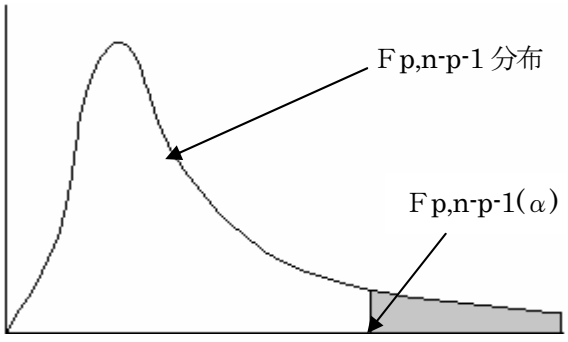
⑴仮説をたてる

仮　　説　$H_0$：$\rho=0$　（母重相関係数は無相関である）

対立仮説　$H_1$：$\rho\neq0$　（母重相関係数は無相関ではない）

⑵検定統計量Ｆは自由度ｐ，ｎ－ｐ－１のＦ分布に従う。

⑶有為水準$\alpha$で検定を実行する。



Ｆ≧Ｆp,n－p－1(α)であれば、仮説を棄却する。つまり、母重相関係数は有効であり、実測値と予測値の間には相関があるといえる。

重相関係数は、実測値ｙと予測値Ｙとの相関係数である。これに対して単純に２変量間の相関係数を単相関係数という。多変量データにおいて、２変量間の相関係数が本当に正しい相関を示すとは限らない。多変量においては２変量間の相関係数を求めても、その２変量以外の変量がこの２変量に影響を与えるからである。よって、多変量間における２変量の正しい相関係数を求めるには、相関係数を求める２変量以外の変量の影響を取り除いて（一定にして）相関係数を求める必要がある。このようにして求めた相関係数を偏相関係数という。

（2）偏相関係数

多変量データにおいて、任意の２変量間の単純な相関係数を単相関係数というが、これは相関をとる２変量以外の変量が、その２変量に影響を与えている相関係数である。これに対し、相関を求める２変量以外の他の変量の影響を取り除いた２変量間の相関係数を偏相関係数という。

いまＰ変量の任意の２変量間の単相関係数を $r_{ij}$ とする。

| | $x_1$ | $x_2$ | … | $x_p$ |
|---|---|---|---|---|
| $x_1$ | $r_{11}$ | $r_{21}$ | … | $r_{p1}$ |
| $x_2$ | $r_{21}$ | $r_{22}$ | … | $r_{p2}$ |
| … | | | … | |
| $x_p$ | $r_{p1}$ | $r_{2p}$ | … | $r_{pp}$ |

単相関行列を$\mathbf{R}$とすると

$$\mathbf{R}=\begin{pmatrix} r_{11} & r_{21} & \cdots & r_{p1} \\ r_{21} & r_{22} & \cdots & r_{p2} \\ & \cdots & & \\ r_{p1} & r_{2p} & \cdots & r_{pp} \end{pmatrix}$$

逆行列$\mathbf{R}^{-1}$を

$$\mathbf{R}^{-1}=\begin{pmatrix} r^{11} & r^{21} & \cdots & r^{p1} \\ r^{21} & r^{22} & \cdots & r^{p2} \\ & \cdots & & \\ r^{p1} & r^{2p} & \cdots & r^{pp} \end{pmatrix}$$

成分 ij 以外の変量を一定にした成分 i･j 間の偏相関係数を $r_{ij.p\cdots q}$ とする。

$$r_{ij.p\cdots q}=\frac{-r^{ij}}{\sqrt{r^{ii}\cdot r^{jj}}}$$

（3）自由度調整済み決定係数

決定係数や重相関係数は、説明変量の数を増やすと単純に増加する傾向がある。

そこで、単純に説明変量の数を増やしても、決定係数が単純に増加しないように調整した自由度調整済み決定係数という。通常標本数がｎ個、説明変量がｎ－１個のものは分析することができない。必ず説明変量がｎ－２個以下にする必要がある。

自由度調整済み決定係数を$R'^2$とすると

$$R'^2 = 1 - \frac{\frac{S_E}{n-p-1}}{\frac{S_T}{n-1}}$$

ｎ：標本数　Ｐ：説明変量の個数（Ｐ＝ｎ－１の時には分母が０になってしまう）

また書き換えると

$$R'^2 = 1 - \frac{n-1}{n-p-1} \cdot (1-R^2)$$

自由度調整済み重相関係数をＲ’とすると　$R' = \sqrt{R'^2}$　である。

１．５　回帰式の信頼性

回帰式を使用して説明変量から目的変量の値を予測する時、その予測値がどのくらい信頼性があるのかを検定する方法に、分散分析を用いる方法と相関係数を用いる方法がある。

1.5.1　分散分析を用いる場合

（１）単回帰のとき

説明変量ｘと実測値ｙと単回帰式から求めた予測値Ｙが下表のようである時

| 標　本 | 説明変量ｘ | 実測値ｙ | 予測値Ｙ |
|---|---|---|---|
| 1 | $x_1$ | $y_1$ | $Y_1$ |
| 2 | $x_2$ | $y_2$ | $Y_2$ |
| … | … | … | … |
| n | $x_n$ | $y_n$ | $Y_n$ |

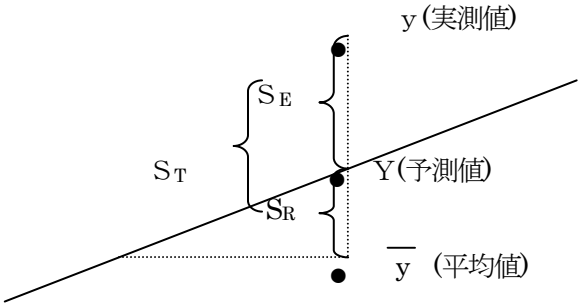
予測値Ｙｉは、$Y = b_1 \cdot x + b_0$の回帰式から求めた値

以上のデータをもとに、分散分析表を作成し回帰式の信頼性を検定する。

全体の変動（$S_T$）を、回帰による変動（$S_R$）と残差による変動（$S_E$）とに分け、回帰による変動が残差による変動よりも小さいようであれば、回帰直線で求めた予測値は残差による影響の方が大きいので予測には役立たないと考える。

実測値の変動（$S_T$）＝回帰による変動（$S_R$）＋残差による変動（$S_E$）

残差が小さいほど「実測値の変動」≒「回帰による変動」となり、よい予測値を得られる。




(1)変動を求める

①実測値の全変動（$S_T$）…実測値の各値$y_i$が、実測値の平均$\overline{y}$からどれ位ばらついているかである。　　$S_T = \Sigma (y_i - \overline{y})^2$

②回帰による変動（$S_R$）…回帰直線によって求めた予測値Yiが、実測値の平均ｙからどれ位ばらついているか である。　$S_R=\Sigma\left(Y_i-\overline{y}\right)^2$

③残差による変動（$S_E$）　$S_E=\Sigma\left(y_i-Y_i\right)^2$

⑵自由度を求める

①回帰による変動の自由度（$f_R$）　$f_R=2-1=1$

②残差による変動の自由度（$f_E$）　$f_E=n-2$

③全変動の自由度（$f_T$）　$f_T=f_R-f_E=n-1$

⑶不偏分散を求める

①回帰による変動の不偏分散（$V_R$）　$V_R=\dfrac{S_R}{F_R}=S_R$

②残差による変動の不偏分散（$V_E$）　$V_E=\dfrac{S_E}{F_E}=\dfrac{S_E}{F_E}$

③全変動の不偏分散（$V_T$）　$V_T=\dfrac{S_T}{F_T}=\dfrac{S_T}{F_T}$

⑷分散比Ｆを求める

$F=\dfrac{V_R}{V_R}=\dfrac{S_R}{S_E/(n-2)}$　は自由度１，ｎ－２のＦ分布に従う。

右片側検定を行い、$V_R$が$V_E$より大きいかどうか検定する。$V_R>V_E$であれば、回帰による変動が残差による変動よりも全変動に与える影響が大きいので､回帰直線は予測に役立つといえる。

⑸検定を行う

⑴仮説をたてる

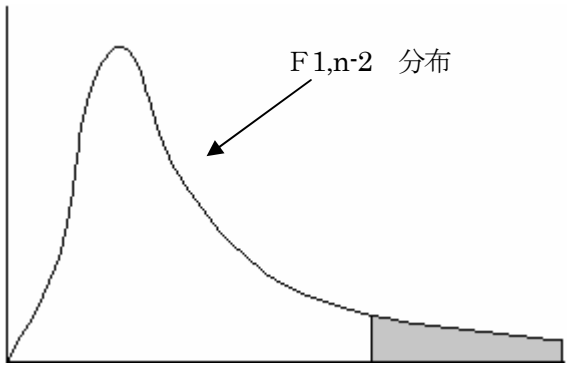
仮　　説　$H_0$：回帰直線は予測に役立たない（$V_R \fallingdotseq V_E$）

対立仮説　$H_1$：回帰直線は予測に役立つ（$V_R>V_E$）

⑵検定統計量Ｆを求める

$F=\dfrac{V_R}{V_R}$　は自由度１，ｎ－２のＦ分布に従う

⑶有為水準αで右片側検定を行う




↑

$F_{1,n-2}(\alpha)$

$F \geqq F_{1,n-2}(\alpha)$であれば、仮説$H_0$を棄却し、対立仮説$H_1$：回帰直線は予測に役立つを採択する。つまり、この回帰直線は予測に役立つとする。