4. 重回帰分析解析

概要: 目的変数は一般には多くの変数の影響を受ける。したがって、一つの説明変数を仮定していた回帰分析を多くの変数を扱う解析へ拡張する。ここでは、多変数の場合の回帰式の導出、回帰式の精度の評価を行う。

キーワード:回帰; 重回帰; 目的変数; 説明変数;重相関係数;梃子比;偏相関係数.

4.1. 序

アパートの値段は、間取り、築年数、駅からの距離などによって決まってくる。つまり、前の章では一つの説明変数と一つの目的変数の関連性を解析してきたが、目的変数に影響を及ぼしているのは一般に一つの変数だけではなく、複数からなる。そこで、このような多くの説明変数と目的変数の関連を解析する必要がある。

この場合、複数の説明変数が独立であれば、先に議論した回帰分析で説明変数の数を増やしたことにしかならない。つまり、基本的には目的変数と一つの説明変数の関係をみればいい。しかし、説明変数が複数になると説明変数間の相互作用が出てくる。それを考慮したものが重回帰分析である。

ここでは、まず二つの説明変数について詳細に解析を行う。その場合、二つの説明変数の相互作用を考慮する場合を考える。それを後に 3 つ以上の多変数解析に拡張する。さらに、問題を行列表現形式にする。行列形式においては、変数の数に依存せず理論の形式は同じであることを示す。

4.2. 二つの説明変数による回帰分析

ここではまず、簡単化のために一つの目的変数yと二つの説明変数x_1およびx_2を扱う。

4.2.1. 二つの説明変数の場合の回帰直線

我々はnセットの標本データを母集団から取り、母集団の特性を予測する。

目的変数y_kと二つの説明変数x_{k1}、x_{k2}は以下のように関連づけられるとする。

$$y_k = a_0 + a_1 x_{k1} + a_2 x_{k2} + e_k \tag{1}$$

ここでnセットの各サンプルを添え字kで表す。a_0，a_1，および a_2 は母集団に関連する定数である。

係数a_0，a_1，およびa_2を標本データから予想する。これらの係数の値は、取得した標本データに依存し、そのたび毎に変動する。つまり、それは母集団のそれとは異なる。したがって、標本データから予想したこれらの係数を$\hat{a}_0$，$\hat{a}_1$，および $\hat{a}_2$ と表記する。

n セットの標本データは

$$\begin{aligned} y_1 &= \hat{a}_0 + \hat{a}_1 x_{11} + \hat{a}_2 x_{12} + e_1 \\ y_2 &= \hat{a}_0 + \hat{a}_1 x_{21} + \hat{a}_2 x_{22} + e_2 \\ &\cdots \\ y_n &= \hat{a}_0 + \hat{a}_1 x_{n1} + \hat{a}_2 x_{n2} + e_n \end{aligned} \tag{2}$$

と表記される。ここで、e_k は残差誤差と呼ばれる。標本回帰値 Y_k は

$$Y_k = \hat{a}_0 + \hat{a}_1 x_{k1} + \hat{a}_2 x_{k2} \tag{3}$$

と表現される。したがって、残差誤差e_kは

$$e_k = y_k - Y_k = y_k - \left(\hat{a}_0 + \hat{a}_1 x_{k1} + \hat{a}_2 x_{k2}\right) \tag{4}$$

となる。

この残差誤差項の平均は 0 で、かつ x_k と独立とする。つまり

$$E[e] = 0 \tag{5}$$

$$Cov[x, e] = 0 \tag{6}$$

とする。これらの誤差項に関連する仮定は妥当なものであるが、その妥当性は次の節で示す。

残差誤差e_kの分散は$S_e^{(2)}$と表記され、

$$S_e^{(2)} = \frac{1}{n}\sum_{k=1}^{n}\left[y_k - \left(\hat{a}_0 + \hat{a}_1 x_{k1} + \hat{a}_2 x_{k2}\right)\right]^2 \tag{7}$$

となる。我々は、この残差誤差分散$S_e^{(2)}$が最小になるように係数 $\hat{a}_0$，$\hat{a}_1$，および $\hat{a}_2$ を以下のように決定する。

$S_e^{(2)}$ を $\hat{a}_0$ に関して偏微分し 0 と置くと

$$\frac{\partial S_e^{(2)}}{\partial \hat{a}_0} = -2\frac{1}{n}\sum_{k=1}^{n}\left[y_k - \left(\hat{a}_0 + \hat{a}_1 x_{k1} + \hat{a}_2 x_{k2}\right)\right] = 0 \tag{8}$$

を得る。これから

$$\begin{aligned} \frac{1}{n}\sum_{k=1}^{n}\left[y_k - \left(\hat{a}_0 + \hat{a}_1 x_{k1} + \hat{a}_2 x_{k2}\right)\right] &= \frac{1}{n}\sum_{k=1}^{n} y_k - \hat{a}_0 - \hat{a}_1 \frac{1}{n}\sum_{i=1}^{n} x_{k1} - \hat{a}_2 \frac{1}{n}\sum_{k=1}^{n} x_{k2} \\ &= \bar{y} - \hat{a}_0 - \hat{a}_1 \bar{x}_1 - \hat{a}_2 \bar{x}_2 \\ &= 0 \end{aligned} \tag{9}$$

を得る。ただし、各変数の平均$\bar{x}_1, \bar{x}_2$、$\bar{y}$は

$$\bar{x}_1 = \frac{1}{n}\sum_{k=1}^{n} x_{k1} \tag{10}$$

$$\bar{x}_2 = \frac{1}{n}\sum_{k=1}^{n} x_{k2} \tag{11}$$

$$\bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k \tag{12}$$

である。

$S_e^{(2)}$ を $\hat{a}_1$ および $\hat{a}_2$ で偏微分し 0 と置くと

$$\frac{\partial S_e^{(2)}}{\partial \hat{a}_1} = -2\frac{1}{n}\sum_{k=1}^{n} x_{k1}\left[y_k - \left(\hat{a}_0 + \hat{a}_1 x_{k1} + \hat{a}_2 x_{k2}\right)\right] = 0 \tag{13}$$

$$\frac{\partial S_e^{(2)}}{\partial \hat{a}_2} = -2\frac{1}{n}\sum_{k=1}^{n} x_{k2}\left[y_k - \left(\hat{a}_0 + \hat{a}_1 x_{k1} + \hat{a}_2 x_{k2}\right)\right] = 0 \tag{14}$$

となる Eq. (9)を整理して再掲すると

$$\bar{y} - \hat{a}_0 - \hat{a}_1\bar{x}_1 - \hat{a}_2\bar{x}_2 = 0 \tag{15}$$

である。これを Eq. (13), (14)に代入して

$$\frac{1}{n}\sum_{k=1}^{n} x_{k1}\left[\left(y_k - \bar{y}\right) - \hat{a}_1\left(x_{k1} - \bar{x}_1\right) - \hat{a}_2\left(x_{k2} - \bar{x}_2\right)\right] = 0 \tag{16}$$

$$\frac{1}{n}\sum_{k=1}^{n} x_{k2}\left[\left(y_k - \bar{y}\right) - \hat{a}_1\left(x_{k1} - \bar{x}_1\right) - \hat{a}_2\left(x_{k2} - \bar{x}_2\right)\right] = 0 \tag{17}$$

を得る。ここで、以下が成り立つことに注目する。

$$\frac{1}{n}\sum_{k=1}^{n} \bar{x}_1\left[\left(y_k - \bar{y}\right) - \hat{a}_1\left(x_{k1} - \bar{x}_1\right) - \hat{a}_2\left(x_{k2} - \bar{x}_2\right)\right] = 0 \tag{18}$$

$$\frac{1}{n}\sum_{k=1}^{n} \bar{x}_2\left[\left(y_k - \bar{y}\right) - \hat{a}_1\left(x_{k1} - \bar{x}_1\right) - \hat{a}_2\left(x_{k2} - \bar{x}_2\right)\right] = 0 \tag{19}$$

これが成り立つことは以下のように証明される。

残差誤差 e_k は

$$e_k = y_k - Y_k = y_k - \left(\hat{a}_0 + \hat{a}_1 x_{k1} + \hat{a}_2 x_{k2}\right) \tag{20}$$

である。この平均が 0 であるから、

$$\frac{1}{n}\sum_{k=1}^{n} e_k = \frac{1}{n}\sum_{k=1}^{n}\left[y_k - \left(\hat{a}_0 + \hat{a}_1 x_{k1} + \hat{a}_2 x_{k2}\right)\right] = 0 \tag{21}$$

である。一方、

$$\bar{y} - \hat{a}_0 - \hat{a}_1\bar{x}_1 - \hat{a}_2\bar{x}_2 = 0 \tag{22}$$

であるから、これを引いて

$$\frac{1}{n}\sum_{k=1}^{n} e_k = \frac{1}{n}\sum_{k=1}^{n}\left[\left(y_k - \bar{y}\right) - \hat{a}_1\left(x_{k1} - \bar{x}_1\right) - \hat{a}_2\left(x_{k2} - \bar{x}_2\right)\right] = 0 \tag{23}$$

となる。$\bar{x}_1, \bar{x}_2$ は k に依存しないから、仮定した二つの式(18), (19)が成り立つ。

これらを利用して

$$
\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n}\left(x_{ki}-\bar{x}_1\right)\left[\left(y_k-\bar{y}\right)-\hat{a}_1\left(x_{k1}-\bar{x}_1\right)-\hat{a}_2\left(x_{k2}-\bar{x}_2\right)\right]\\
&=\frac{1}{n}\sum_{k=1}^{n}\left(x_{ki}-\bar{x}_1\right)\left(y_k-\bar{y}\right)-\hat{a}_1\frac{1}{n}\sum_{k=1}^{n}\left(x_{ki}-\bar{x}_1\right)\left(x_{k1}-\bar{x}_1\right)-\hat{a}_2\frac{1}{n}\sum_{k=1}^{n}\left(x_{ki}-\bar{x}_1\right)\left(x_{k2}-\bar{x}_2\right)\\
&=S_{1y}^{(2)}-\hat{a}_1S_{11}^{(2)}-\hat{a}_2S_{12}^{(2)}\\
&=0
\end{aligned}
\tag{24}
$$

$$
\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n}\left(x_{k2}-\bar{x}_2\right)\left[\left(y_k-\bar{y}\right)-\hat{a}_1\left(x_{k1}-\bar{x}_1\right)-\hat{a}_2\left(x_{k2}-\bar{x}_2\right)\right]\\
&=\frac{1}{n}\sum_{k=1}^{n}\left(x_{k2}-\bar{x}_2\right)\left(y_k-\bar{y}\right)-\hat{a}_1\frac{1}{n}\sum_{k=1}^{n}\left(x_{k2}-\bar{x}_2\right)\left(x_{k1}-\bar{x}_1\right)-\hat{a}_2\frac{1}{n}\sum_{k=1}^{n}\left(x_{k2}-\bar{x}_2\right)\left(x_{k2}-\bar{x}_2\right)\\
&=S_{2y}^{(2)}-\hat{a}_1S_{21}^{(2)}-\hat{a}_2S_{22}^{(2)}\\
&=0
\end{aligned}
\tag{25}
$$

となる。つまり

$$\hat{a}_1S_{11}^{(2)}+\hat{a}_2S_{12}^{(2)}=S_{1y}^{(2)} \tag{26}$$

$$\hat{a}_1S_{21}^{(2)}+\hat{a}_2S_{22}^{(2)}=S_{2y}^{(2)} \tag{27}$$

を得る。ただし、以下の分散を定義している。

$$S_{11}^{(2)}\equiv\frac{1}{n}\sum_{k=1}^{n}\left(x_{k1}-\bar{x}_1\right)^2 \tag{28}$$

$$S_{22}^{(2)}\equiv\frac{1}{n}\sum_{k=1}^{n}\left(x_{k2}-\bar{x}_2\right)^2 \tag{29}$$

$$
\begin{aligned}
S_{12}^{(2)}&=S_{21}^{(2)}\\
&=\frac{1}{n}\sum_{k=1}^{n}\left(x_{k1}-\bar{x}_1\right)\left(x_{k2}-\bar{x}_2\right)
\end{aligned}
\tag{30}
$$

$$S_{yy}^{(2)}\equiv\frac{1}{n}\sum_{k=1}^{n}\left(y_k-\bar{y}\right)^2 \tag{31}$$

$$S_{1y}^{(2)}\equiv\frac{1}{n}\sum_{k=1}^{n}\left(x_{k1}-\bar{x}_1\right)\left(y_k-\bar{y}\right) \tag{32}$$

$$S_{2y}^{(2)}=\frac{1}{n}\sum_{k=1}^{n}\left(x_{k2}-\bar{x}_2\right)\left(y_k-\bar{y}\right) \tag{33}$$

この関係式を行列表現すると以下となる。

$$\begin{pmatrix}S_{11}^{(2)} & S_{12}^{(2)}\\ S_{21}^{(2)} & S_{22}^{(2)}\end{pmatrix}\begin{pmatrix}\hat{a}_1\\ \hat{a}_2\end{pmatrix}=\begin{pmatrix}S_{1y}^{(2)}\\ S_{2y}^{(2)}\end{pmatrix} \tag{34}$$

これから $\hat{a}_1$ および $\hat{a}_2$ は以下のように求めることができる。

$$\begin{aligned}\begin{pmatrix}\hat{a}_1\\ \hat{a}_2\end{pmatrix}&=\begin{pmatrix}S_{11}^{(2)} & S_{12}^{(2)}\\ S_{12}^{(2)} & S_{22}^{(2)}\end{pmatrix}^{-1}\begin{pmatrix}S_{1y}^{(2)}\\ S_{2y}^{(2)}\end{pmatrix}\\ &=\begin{pmatrix}S^{11(2)} & S^{12(2)}\\ S^{21(2)} & S^{22(2)}\end{pmatrix}\begin{pmatrix}S_{1y}^{(2)}\\ S_{2y}^{(2)}\end{pmatrix}\end{aligned} \tag{35}$$

ただし、

$$\begin{pmatrix}S^{11(2)} & S^{12(2)}\\ S^{21(2)} & S^{22(2)}\end{pmatrix}=\begin{pmatrix}S_{11}^{(2)} & S_{12}^{(2)}\\ S_{12}^{(2)} & S_{22}^{(2)}\end{pmatrix}^{-1} \tag{36}$$

である。2 行 2 列の行列の場合は

$$\begin{aligned}\begin{pmatrix}S^{11(2)} & S^{12(2)}\\ S^{21(2)} & S^{22(2)}\end{pmatrix}&=\begin{pmatrix}S_{11}^{(2)} & S_{12}^{(2)}\\ S_{12}^{(2)} & S_{22}^{(2)}\end{pmatrix}^{-1}\\ &=\frac{1}{S_{11}^{(2)}S_{22}^{(2)}-S_{12}^{(2)}S_{12}^{(2)}}\begin{pmatrix}S_{22}^{(2)} & -S_{12}^{(2)}\\ -S_{12}^{(2)} & S_{11}^{(2)}\end{pmatrix}\end{aligned} \tag{37}$$

と逆行列は解析的に求まる。つまり、

$$\begin{aligned}\begin{pmatrix}\hat{a}_1\\ \hat{a}_2\end{pmatrix}&=\begin{pmatrix}S^{11(2)} & S^{12(2)}\\ S^{21(2)} & S^{22(2)}\end{pmatrix}\begin{pmatrix}S_{1y}^{(2)}\\ S_{2y}^{(2)}\end{pmatrix}\\ &=\frac{1}{S_{11}^{(2)}S_{22}^{(2)}-S_{12}^{(2)}S_{12}^{(2)}}\begin{pmatrix}S_{22}^{(2)} & -S_{12}^{(2)}\\ -S_{12}^{(2)} & S_{11}^{(2)}\end{pmatrix}\begin{pmatrix}S_{1y}^{(2)}\\ S_{2y}^{(2)}\end{pmatrix}\\ &=\frac{1}{S_{11}^{(2)}S_{22}^{(2)}-S_{12}^{(2)}S_{12}^{(2)}}\begin{pmatrix}S_{22}^{(2)}S_{1y}^{(2)}-S_{12}^{(2)}S_{2y}^{(2)}\\ -S_{12}^{(2)}S_{1y}^{(2)}+S_{11}^{(2)}S_{2y}^{(2)}\end{pmatrix}\end{aligned} \tag{38}$$

となる。2 次より多い次数の一般の場合は数値的に解いて求める。

ここで、$\hat{a}_1$ および $\hat{a}_2$ が Eq. (38)から求まると，$\hat{a}_0$ は Eq. (9)から

$$\hat{a}_0=\bar{y}-\hat{a}_1\bar{x}_1-\hat{a}_2\bar{x}_2 \tag{39}$$

と求まる。

説明変数同士が独立であると仮定すると以下のようになる。

説明変数が独立の場合は

$$\begin{cases}S_{12}^{(2)}\approx 0\\ S_{21}^{(2)}\approx 0\end{cases} \tag{40}$$

と近似できる。したがって、我々は$\hat{a}_1$および$\hat{a}_2$ を

$$
\begin{aligned}
\begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \end{pmatrix} &= \begin{pmatrix} S_{11}^{(2)} & 0 \\ 0 & S_{22}^{(2)} \end{pmatrix}^{-1} \begin{pmatrix} S_{1y}^{(2)} \\ S_{2y}^{(2)} \end{pmatrix} \\
&= \frac{1}{S_{11}^{(2)} S_{22}^{(2)}} \begin{pmatrix} S_{22}^{(2)} & 0 \\ 0 & S_{11}^{(2)} \end{pmatrix} \begin{pmatrix} S_{1y}^{(2)} \\ S_{2y}^{(2)} \end{pmatrix} \\
&= \begin{pmatrix} \dfrac{S_{1y}^{(2)}}{S_{11}^{(2)}} \\ \dfrac{S_{2y}^{(2)}}{S_{22}^{(2)}} \end{pmatrix}
\end{aligned}
\tag{41}
$$

と求めることができる。ただし、

ここで、$\hat{a}_1$ および $\hat{a}_2$ が Eq. (41)から求まると，$\hat{a}_0$ は Eq. (9)から

$$
\begin{aligned}
\hat{a}_0 &= \bar{y} - \hat{a}_1 \bar{x}_1 - \hat{a}_2 \bar{x}_2 \\
&= \bar{y} - \frac{S_{1y}^{(2)}}{S_{11}^{(2)}{}_1} \bar{x}_1 - \frac{S_{2y}^{(2)}}{S_{22}^{(2)}} \bar{x}_2
\end{aligned}
\tag{42}
$$

と求まる。つまり、回帰直線は

$$
\begin{aligned}
Y_k &= \hat{a}_0 + \hat{a}_1 x_{k1} + \hat{a}_2 x_{k2} \\
&= \bar{y} - \frac{S_{1y}^{(2)}}{S_{11}^{(2)}} \bar{x}_1 - \frac{S_{2y}^{(2)}}{S_{22}^{(2)}} \bar{x}_2 + \frac{S_{1y}^{(2)}}{S_{11}^{(2)}} x_{k1} + \frac{S_{2y}^{(2)}}{S_{22}^{(2)}} x_{k2}
\end{aligned}
\tag{43}
$$

となる。整理して

$$
Y_k - \bar{y} = \frac{S_{1y}^{(2)}}{S_{11}^{(2)}} \left(x_{k1} - \bar{x}_1 \right) + \frac{S_{2y}^{(2)}}{S_{22}^{(2)}} \left(x_{k2} - \bar{x}_2 \right)
\tag{44}
$$

さらに両辺を $S_{yy}^{(2)}$ で割ると

$$
\begin{aligned}
\frac{Y_k - \bar{y}}{\sqrt{S_{yy}^{(2)}}} &= \frac{S_{1y}^{(2)}}{\sqrt{S_{11}^{(2)} S_{yy}^{(2)}}} \left(\frac{x_{k1} - \bar{x}_1}{\sqrt{S_{11}^{(2)}}} \right) + \frac{S_{2y}^{(2)}}{\sqrt{S_{22}^{(2)} S_{yy}^{(2)}}} \left(\frac{x_{k2} - \bar{x}_2}{\sqrt{S_{22}^{(2)}}} \right) \\
&= r_{1y} \left(\frac{x_{k1} - \bar{x}_1}{\sqrt{S_{11}^{(2)}}} \right) + r_{2y} \left(\frac{x_{k2} - \bar{x}_2}{\sqrt{S_{22}^{(2)}}} \right)
\end{aligned}
\tag{45}
$$

となる。これは、前章で求めた 1 変数の場合の回帰式、

$$
\frac{Y_k - \bar{y}}{\sqrt{S_{yy}^{(2)}}} = r_{1y} \left(\frac{x_{k1} - \bar{x}_1}{\sqrt{S_{11}^{(2)}}} \right)
\tag{46}
$$

に項を一つ足した場合に相当する。

4.2.2. 残差誤差e_iの評価

ここでは、証明なしに仮定していた誤差誤差e_kの特性に関して検討する。

e_kの平均は$\bar{e}$と表記する。これは

$$\begin{aligned}\bar{e} &= \frac{1}{n}\sum_{k=1}^{n} e_k \\ &= \frac{1}{n}\sum_{k=1}^{n}\left(y_k - Y_k\right) \\ &= \frac{1}{n}\sum_{k=1}^{n}\left[y_k - \left(\hat{a}_0 + \hat{a}_1 x_{k1} + \hat{a}_2 x_{k2}\right)\right] \\ &= \frac{1}{n}\sum_{k=1}^{n}\left[\left(y_k - \bar{y}\right) - \hat{a}_1\left(x_{k1} - \bar{x}_1\right) - \hat{a}_2\left(x_{k2} - \bar{x}_2\right)\right] \\ &= 0\end{aligned} \tag{47}$$

となる。つまり、平均は、仮定していた通りに0になる。

これとx_1の共分散は

$$\begin{aligned}Cov[x_1, e] &= \frac{1}{n}\sum_{k=1}^{n}\left(x_{k1} - \bar{x}_1\right)e_k \\ &= \frac{1}{n}\sum_{k=1}^{n}\left(x_{k1} - \bar{x}_1\right)\left[y_k - \left(\hat{a}_0 + \hat{a}_1 x_k + \hat{a}_2 x_{k2}\right)\right] \\ &= \frac{1}{n}\sum_{k=1}^{n}\left(x_{k1} - \bar{x}_1\right)\left[\left(y_k - \bar{y}\right) - \hat{a}_1\left(x_{k1} - \bar{x}_1\right) - \hat{a}_2\left(x_{k2} - \bar{x}_2\right)\right] \\ &= S_{1y}^{(2)} - \hat{a}_1 S_{11}^{(2)} - \hat{a}_2 S_{12}^{(2)} \\ &= S_{1y}^{(2)} - S_{1y}^{(2)} \\ &= 0\end{aligned} \tag{48}$$

したがって、誤差とx_1との共分散も仮定していた通り0になる。

これとx_2の共分散は

$$\begin{aligned}Cov[x_2, e] &= \frac{1}{n}\sum_{k=1}^{n}\left(x_{k2} - \bar{x}_2\right)e_k \\ &= \frac{1}{n}\sum_{k=1}^{n}\left(x_{k2} - \bar{x}_2\right)\left[y_k - \left(\hat{a}_0 + \hat{a}_1 x_{k1} + \hat{a}_2 x_{k2}\right)\right] \\ &= \frac{1}{n}\sum_{k=1}^{n}\left(x_{k2} - \bar{x}_2\right)\left[\left(y_k - \bar{y}\right) - \hat{a}_1\left(x_{k1} - \bar{x}_1\right) - \hat{a}_2\left(x_{k2} - \bar{x}_2\right)\right] \\ &= S_{2y}^{(2)} - \hat{a}_1 S_{21}^{(2)} - \hat{a}_2 S_{22}^{(2)} \\ &= S_{2y}^{(2)} - S_{2y}^{(2)} \\ &= 0\end{aligned} \tag{49}$$

したがって、誤差とx_2との共分散も仮定していた通り0になる。

e_kの分散は

$$
\begin{aligned}
S_e^{(2)} &= \frac{e_1^{\,2} + e_2^{\,2} + \cdots + e_n^{\,2}}{n} \\
&= \frac{1}{n}\sum_{k=1}^{n}\left(y_k - \hat{a}_0 - \hat{a}_1 x_{k1} - \hat{a}_2 x_{k2}\right)^2 \\
&= \frac{1}{n}\sum_{k=1}^{n}\left[\left(y_k - \bar{y}\right) - \hat{a}_1\left(x_{k1} - \bar{x}_1\right) - \hat{a}_2\left(x_{k2} - \bar{x}_2\right)\right]^2 \\
&= \frac{1}{n}\sum_{k=1}^{n}\left(y_k - \bar{y}\right)^2 + \hat{a}_1^{\,2}\frac{1}{n}\sum_{k=1}^{n}\left(x_{k1} - \bar{x}_1\right)^2 + \hat{a}_2^{\,2}\frac{1}{n}\sum_{k=1}^{n}\left(x_{k2} - \bar{x}_2\right)^2 \\
&\quad -2\hat{a}_1\frac{1}{n}\sum_{k=1}^{n}\left(y_k - \bar{y}\right)\left(x_{k1} - \bar{x}_1\right) - 2\hat{a}_2\frac{1}{n}\sum_{k=1}^{n}\left(y_k - \bar{y}\right)\left(x_{k2} - \bar{x}_2\right) \\
&\quad +2\hat{a}_1\hat{a}_2\frac{1}{n}\sum_{k=1}^{n}\left(x_{k1} - \bar{x}_1\right)\left(x_{k2} - \bar{x}_2\right) \\
&= S_{yy}^{(2)} + \hat{a}_1^{\,2}S_{11}^{(2)} + \hat{a}_2^{\,2}S_{22}^{(2)} - 2\hat{a}_1 S_{1y}^{(2)} - 2\hat{a}_2 S_{2y}^{(2)} + 2\hat{a}_1\hat{a}_2 S_{12}^{(2)} \\
&= S_{yy}^{(2)} + \hat{a}_1\left(S_{11}^{(2)} + \hat{a}_2 S_{12}^{(2)}\right) + \hat{a}_2\left(\hat{a}_2 S_{22}^{(2)} + \hat{a}_1 S_{12}^{(2)}\right) - 2\hat{a}_1 S_{1y}^{(2)} - 2\hat{a}_2 S_{2y}^{(2)} \\
&= S_{yy}^{(2)} - \hat{a}_1 S_{1y}^{(2)} - \hat{a}_2 S_{2y}^{(2)}
\end{aligned}
\tag{50}
$$

となる。これは、0 になる必要はない。つまり、0 になることを仮定していない。

以上より、残差誤差分散 $S_e^{(2)}$ を最小になるように回帰直線の係数を定めると、残差誤差に対する仮定である平均は 0 で他の変数との共分散は 0 であるという仮定は自動的に成り立っている。

4.2.3. 回帰分散による回帰直線の精度の評価

ここでは、回帰直線の精度を議論する。回帰直線の値と実際の目的変数の値の差の分散は $S_e^{(2)}$ であり、これを最小にするように係数が定められた。したがって、回帰直線の精度は $S_e^{(2)}$ と関連づける。一方、目的変数の分散 $S_{yy}^{(2)}$ はこの残差誤差の分散 $S_e^{(2)}$ と回帰直線の分散 $S_R^{(2)}$ の和からなることが後に示される。したがって、この $S_R^{(2)}$ の $S_{yy}^{(2)}$ に対する比を回帰直線の精度の指標にできる。つまり

$$
R^2 = \frac{S_R^{(2)}}{S_{yy}^{(2)}} \tag{51}
$$

を精度の指標にする。この R^2 は決定係数と呼ばれる。ただし $S_R^{(2)}$ は回帰直線上の点 Y_k と関連する分散であり

$$
S_R^{(2)} = \frac{1}{n}\sum_{k=1}^{n}\left(Y_k - \bar{y}\right)^2 \tag{52}
$$

である。誤差が 0、すなわち点が回帰線上に全てのっていれば、R^2 は 1 になる。つまり、

R^2 は 0 から 1 の間の値を取り、1 に近いほど回帰の精度が高いと言える。一方、回帰直線のデータと実際の目的変数の値の差が大きければ、 $S_R^{(2)}$ に比べて $S_e^{(2)}$ が大きくなる。つまり $S_{yy}^{(2)}$ が大きくなり決定係数 R^2 は 1 より大分小さくなる。

上で利用した分散の関係を導く。

y の分散は以下のように導かれる。

$$
\begin{aligned}
S_{yy}^{(2)} &= \frac{1}{n}\sum_{k=1}^{n}\left(y_k-\bar{y}\right)^2 \\
&= \frac{1}{n}\sum_{k=1}^{n}\left(y_k-Y_k+Y_k-\bar{y}\right)^2 \\
&= \frac{1}{n}\sum_{k=1}^{n}\left(y_k-Y_k\right)^2+\frac{1}{n}\sum_{k=1}^{n}\left(Y_k-\bar{y}\right)^2+2\frac{1}{n}\sum_{k=1}^{n}\left(y_k-Y_k\right)\left(Y_k-\bar{y}\right) \\
&= \frac{1}{n}\sum_{k=1}^{n}e_k^2+\frac{1}{n}\sum_{k=1}^{n}\left(Y_k-\bar{y}\right)^2+2\frac{1}{n}\sum_{k=1}^{n}e_k\left(\hat{a}_0+\hat{a}_1x_{k1}+\hat{a}_2x_{k2}\right) \\
&= \frac{1}{n}\sum_{k=1}^{n}e_k^2+\frac{1}{n}\sum_{k=1}^{n}\left(Y_k-\bar{y}\right)^2+2\hat{a}_0\frac{1}{n}\sum_{k=1}^{n}e_k+2\hat{a}_1\frac{1}{n}\sum_{k=1}^{n}e_kx_{k1}+2\hat{a}_2\frac{1}{n}\sum_{k=1}^{n}e_kx_{k2} \\
&= \frac{1}{n}\sum_{k=1}^{n}e_k^2+\frac{1}{n}\sum_{k=1}^{n}\left(Y_k-\bar{y}\right)^2 \\
&= S_e^{(2)}+S_R^{(2)}
\end{aligned}
\tag{53}
$$

ここで

$$
\frac{1}{n}\sum_{k=1}^{n}e_k=\frac{1}{n}\sum_{k=1}^{n}e_kx_{k1}=\frac{1}{n}\sum_{k=1}^{n}e_kx_{k2}=0 \tag{54}
$$

を利用している。つまり、目的変数 y の分散は残差誤差の分散と回帰分散の和ということになる。

$S_e^{(2)}$ は以下のように変形される。

$$
\begin{aligned}
S_e^{(2)} &= \frac{1}{n}\sum_{k=1}^{n}\left(y_k - Y_k\right)^2 \\
&= \frac{1}{n}\sum_{k=1}^{n}\left[y_k - \left(\hat{a}_0 + \hat{a}_1 x_{k1} + \hat{a}_2 x_{k2}\right)\right]^2 \\
&= \frac{1}{n}\sum_{k=1}^{n}\left\{\left[y_k - \left(\hat{a}_0 + \hat{a}_1 x_{k1} + \hat{a}_2 x_{k2}\right)\right] - \left[\bar{y} - \left(\hat{a}_0 + \hat{a}_1 \bar{x}_1 + \hat{a}_2 \bar{x}_2\right)\right]\right\}^2 \\
&= \frac{1}{n}\sum_{k=1}^{n}\left[\left(y_k - \bar{y}\right) - \hat{a}_1\left(x_{k1} - \bar{x}_1\right) - \hat{a}_2\left(x_{k2} - \bar{x}_2\right)\right]^2 \\
&= \frac{1}{n}\sum_{k=1}^{n}\left(y_k - \bar{y}\right)^2 + \hat{a}_1{}^2\frac{1}{n}\sum_{k=1}^{n}\left(x_{k1} - \bar{x}_1\right)^2 + \hat{a}_2\frac{1}{n}\sum_{k=1}^{n}\left(x_{k2} - \bar{x}_2\right)^2 \\
&\quad + 2\hat{a}_1\hat{a}_2\frac{1}{n}\sum_{k=1}^{n}\left(x_{k1} - \bar{x}_1\right)\left(x_{k2} - \bar{x}_2\right)^2 - 2\hat{a}_1\frac{1}{n}\sum_{k=1}^{n}\left(y_k - \bar{y}\right)\left(x_{k1} - \bar{x}_1\right) - 2\hat{a}_2\frac{1}{n}\sum_{k=1}^{n}\left(y_k - \bar{y}\right)\left(x_{k2} - \bar{x}_2\right) \\
&= S_{yy}^{(2)} + \hat{a}_1{}^2 S_{11}^{(2)} + \hat{a}_2{}^2 S_{22}^{(2)} + 2\hat{a}_1\hat{a}_2 S_{12}^{(2)} - 2\hat{a}_1 S_{1y}^{(2)} - 2\hat{a}_2 S_{2y}^{(2)} \\
&= S_{yy}^{(2)} + \hat{a}_1\left(\hat{a}_1 S_{11}^{(2)} + \hat{a}_2 S_{12}^{(2)}\right) + \hat{a}_2\left(\hat{a}_2 S_{22}^{(2)} + \hat{a}_1 S_{12}^{(2)}\right) - 2\hat{a}_1 S_{1y}^{(2)} - 2\hat{a}_2 S_{2y}^{(2)} \\
&= S_{yy}^{(2)} - \left(\hat{a}_1 S_{1y}^{(2)} + \hat{a}_2 S_{2y}^{(2)}\right)
\end{aligned}
\tag{55}
$$

ここで、先の関係式

$$
\begin{cases}
\hat{a}_1 S_{11}^{(2)} + \hat{a}_2 S_{12}^{(2)} = S_{1y}^{(2)} \\
\hat{a}_1 S_{21}^{(2)} + \hat{a}_2 S_{22}^{(2)} = S_{2y}^{(2)}
\end{cases}
\tag{56}
$$

を利用している。

二つの関係式を比較すると

$$
\begin{aligned}
S_R^{(2)} &= \frac{1}{n}\sum_{k=1}^{n}\left(Y_k - \bar{y}\right)^2 \\
&= \hat{a}_1 S_{1y}^{(2)} + \hat{a}_2 S_{2y}^{(2)} \\
&= \sum_{i=1}^{2}\hat{a}_i S_{iy}^{(2)}
\end{aligned}
\tag{57}
$$

となる。

以上より、決定係数は

$$
\begin{aligned}
R^2 &= \frac{S_R^{(2)}}{S_{yy}^{(2)}} \\
&= \frac{S_{yy}{}^{(2)} - S_e^{(2)}}{S_{yy}^{(2)}} \\
&= 1 - \frac{S_e{}^{(2)}}{S_{yy}{}^{(2)}}
\end{aligned}
\tag{58}
$$

となる。このように決定係数は回帰の精度に応じて 0 から 1 の間の値を取ることが再び示された。精度が高いほど 1 に近づく。

4.2.4. 重相関係数による回帰直線の精度の評価

ここでは、回帰直線の精度を重相関係数から議論する。もし、回帰がうまくいっていれば、回帰データと実際の目的変数のデータの相関係数は 1 に近くなるはずである。したがって、この二つの変数 y_i と Y_i の相関係数を評価すればいい。この回帰相関係数を r_{mult} を置くと、それは

$$r_{mult} = \frac{\sum_{k=1}^{n}\left(y_k - \bar{y}\right)\left(Y_k - \bar{y}\right)}{\sqrt{\sum_{k=1}^{n}\left(y_k - \bar{y}\right)^2}\sqrt{\sum_{k=1}^{n}\left(Y_k - \bar{y}\right)^2}} \tag{59}$$

で与えられる。

この分子を評価する。まず、残差誤差に関する分散を変形すると

$$\begin{aligned} S_e^{(2)} &= \frac{1}{n}\sum_{k=1}^{n}\left(y_k - Y_k\right)^2 \\ &= \frac{1}{n}\sum_{k=1}^{n}\left[\left(y_k - \bar{y}\right) - \left(Y_k - \bar{y}\right)\right]^2 \\ &= \frac{1}{n}\sum_{k=1}^{n}\left(y_k - \bar{y}\right)^2 + \frac{1}{n}\sum_{k=1}^{n}\left(Y_k - \bar{y}\right)^2 - 2\frac{1}{n}\sum_{k=1}^{n}\left(y_k - \bar{y}\right)\left(Y_k - \bar{y}\right) \\ &= S_{yy}^{(2)} + S_R^{(2)} - 2\frac{1}{n}\sum_{k=1}^{n}\left(y_k - \bar{y}\right)\left(Y_k - \bar{y}\right) \end{aligned} \tag{60}$$

となり、重相関係数の分子の項が出てくる。これは

$$\begin{aligned} \frac{1}{n}\sum_{k=1}^{n}\left(y_k - \bar{y}\right)\left(Y_k - \bar{y}\right) &= \frac{1}{2}\left(S_{yy}^{(2)} + S_R^{(2)} - S_e^{(2)}\right) \\ &= \frac{1}{2}\left[S_{yy}^{(2)} + S_R^{(2)} - \left(S_{yy}^{(2)} - S_R^{(2)}\right)\right] \\ &= S_R^{(2)} \end{aligned} \tag{61}$$

となる。したがって、重相関係数は

$$\begin{aligned} r_{mult} &= \frac{S_R^{(2)}}{\sqrt{S_{yy}^{(2)}}\sqrt{S_R^{(2)}}} \\ &= \sqrt{\frac{S_R^{(2)}}{S_{yy}^{(2)}}} \end{aligned} \tag{62}$$

となる。これからすると、重相関係数 r_{mult} の 2 乗が決定係数となる。つまり、重相関係数と決定係数はほぼ同じことをしている。式で表すと

$$r_{mult} = \sqrt{R^2} \tag{63}$$

となる。または、

$$R^2 = r_{mult}{}^2 \tag{64}$$

である。

この重相関係数と先にのべた決定係数は同じく回帰の精度をあらわしている。決定係数は次にのべる自由度を考慮したものに発展させられるが、重相関係数はこのままである。したがって、回帰の精度は次の節で述べる自由度調整済み決定係数で評価されることが多い。

4.2.5. 自由度調整済決定係数

回帰の精度はこれまで示してきたように決定係数または重相関係数で評価できる。この両者はほぼ同じことを評価している。これらは、回帰の精度を記述しているが以下の点で一般性に欠ける。

決定係数は標本の数、変数の数に依存する。これらが異なるもの同士を比べるには決定係数は一般性がない。そこで、この決定係数のさらに精度を上げ、異なる標本数、変数の数に対しても評価できる一般的なものにするる。

決定係数は回帰分散と全体のデータの分散の比であるが、変形すると

$$R^2 = \frac{S_R^{(2)}}{S_{yy}^{(2)}} = \frac{S_{yy}^{(2)} - S_e^{(2)}}{S_{yy}^{(2)}} = 1 - \frac{S_e^{(2)}}{S_{yy}^{(2)}} \tag{65}$$

となる。つまり、同じことであるが、誤差の分散と目的変数の分散の比が小さいと精度は高い、と評価される。しかし、この分散では自由度を考慮していない。ここでは、それらの分散の自由度を評価し、それを評価に取り入れる。

まず、$S_{yy}^{(2)}$ の自由度 ϕ_T を考える。これは n 個のデータからなり、分散を評価するうえでその平均を利用している。つまり、その自由度は

$$\phi_T = n - 1 \tag{66}$$

である。

次に回帰分散 $S_R^{(2)}$ の自由度 ϕ_R を考える。これは 2 個の変数 $\hat{a}_1, \hat{a}_2$ で評価される。つまり、その自由度は

$$\phi_R = p = 2 \tag{67}$$

である。つまり、その自由度は説明変数の数である。

最後に残差誤差分散 $S_e^{(2)}$ の自由度 ϕ_e を考える。これは n 個の目的変数データと、3 個の係数で定まる回帰式の差である。つまり、その自由度は

$$\phi_e = n-(p+1) = n-(2+1) \tag{68}$$

である。つまり、それはデータ数から説明変数と目的変数の数の和を引いたものである。

各分散の関係は

$$S_{yy}^{(2)} = S_R^{(2)} + S_e^{(2)} \tag{69}$$

であり、これは自由度の関係の方程式と直接関連する。すなわち自由度に関して方程式な成り立ち

$$\phi_T = \phi_R + \phi_e \tag{70}$$

である。この関係は常に成り立つ。すなわち、変数の方程式と変数の自由度の方程式は常に 1 対 1 に対応する。

したがって、決定係数はその自由度を考慮した

$$R^{*2} = 1 - \frac{\frac{n}{\phi_e} S_e^{(2)}}{\frac{n}{\phi_T} S_{yy}^{(2)}} \tag{71}$$

で改良され評価される。こちらのほうが、より定量的に意味のある回帰の評価式である。これを調整済決定係数と呼ぶ。一般に、決定係数よりも調整済決定係数のほうが小さい値になる。これに実際の値を入れると

$$\begin{aligned} R^{*2} &= 1 - \frac{\frac{n}{\phi_e} S_e^{(2)}}{\frac{n}{\phi_T} S_{yy}^{(2)}} \\ &= 1 - \frac{\frac{n}{n-3} S_e^{(2)}}{\frac{n}{n-1} S_{yy}^{(2)}} \\ &= 1 - \frac{n-1}{n-3} \frac{S_e^{(2)}}{S_{yy}^{(2)}} \\ &= 1 - \frac{1}{1-\frac{2}{n-1}} \frac{S_e^{(2)}}{S_{yy}^{(2)}} \end{aligned} \tag{72}$$

となり、決定係数よりも小さくなる。

4.2.6. F 値による回帰精度の評価

回帰の精度は決定係数で評価した。それは、全体の分散と回帰の分散の比であり、回帰の精度が高ければ、それは 1 に近づくというものであった。

この回帰の精度は別の観点からも評価できる。これは、回帰分散と残差誤差分散の比か

ら決定できる。この比、

$$F = \frac{s_R^{(2)}}{s_e^{(2)}} \tag{73}$$

が大きければ、回帰の精度はいいと言える。ただし、

$$s_R^{(2)} = \frac{nS_R^{(2)}}{\phi_R} = \frac{nS_R^{(2)}}{2} \tag{74}$$

$$s_e^{(2)} = \frac{nS_e^{(2)}}{\phi_e} = \frac{nS_e^{(2)}}{n-(2+1)} \tag{75}$$

である。ここで各分散の自由度

$$S_R^{(2)} : \phi_R = p = 2 \tag{76}$$

$$S_e^{(2)} : \phi_e = n-(p+1) = n-(2+1) \tag{77}$$

を利用している。これと F 値の推定確率 P の場合の臨界値 $F_p(\phi_R,\phi_e) = F_p(2, n-(2+1))$ と比較し、この F 値が $F_p(2, n-(2+1))$ よりも大きければ回帰の精度は高い、と判断できる。つまり以下の式で表現される。

$$\begin{cases} F \le F_p(2, n-(2+1)) \Rightarrow invalid \\ F > F_p(2, n-(2+1)) \Rightarrow valid \end{cases} \tag{78}$$

一般に、回帰の精度は上の調整済み決定係数か、F 値から判断される。

先に示した調整済み決定係数とこの F 値の評価は同じことを別の観点から評価している。したがって、回帰の精度は調整済み決定係数かこの、F 値どちらかを採用すればいい。F 値を用いるほうが、推定確率と連動させて回帰の精度が十分かそうでないかを判断できる。

4.2.7. 回帰直線の係数の変動

標本データを変えるたびに、取得するデータは異なり、それから導き出す回帰直線も異なったものになる。つまり、回帰直線の係数 $\hat{a}_0$、$\hat{a}_1$、$\hat{a}_2$ は抽出されたデータに応じて変動する。ここでは、$\hat{a}_0$ と $\hat{a}_1$ および $\hat{a}_2$ の平均と分散、共分散を検討する。

この評価において、係数が 0 であれば、その変数は考える必要はないといえる。つまり、係数 $\hat{a}_0$、$\hat{a}_1$、$\hat{a}_2$ が 0 とみなすことができるのかを判定すればいい。

我々は母集団の回帰係数を仮定して、目的変数は以下のように表現されるとする。

$$y_k = a_0 + a_1 x_{k1} + a_2 x_{k2} + e_k \quad for\ k = 1, 2, \cdots, n \tag{79}$$

ここで (x_{k1},x_{k2}) は与えられたデータ，係数 a_0，a_1，および a_2 は確立された値とする。e_i は正規分布 $N\left[0,\sigma_e^{(2)}\right]$ に従う。式の中で確率変数は e_i だけである。

$k=1,2,\cdots,n$ についての n セットのデータ (x_{k1},x_{k2},y_k) を取得したとする。その係数は

$$\begin{aligned}\begin{pmatrix}\hat{a}_1\\ \hat{a}_2\end{pmatrix}&=\begin{pmatrix}S^{11(2)} & S^{12(2)}\\ S^{21(2)} & S^{22(2)}\end{pmatrix}\begin{pmatrix}S_{1y}^{(2)}\\ S_{2y}^{(2)}\end{pmatrix}\\&=\begin{pmatrix}S^{11(2)}S_{1y}^{(2)}+S^{12(2)}S_{2y}^{(2)}\\ S^{21(2)}S_{1y}^{(2)}+S^{22(2)}S_{2y}^{(2)}\end{pmatrix}\end{aligned}\tag{80}$$

となる。

まず、$\hat{a}_1$ を考える。それは、式を展開して

$$\begin{aligned}\hat{a}_1&=S^{11(2)}S_{1y}^{(2)}+S^{12(2)}S_{2y}^{(2)}\\&=\sum_{p=1}^{2}S^{1p(2)}S_{py}^{(2)}\end{aligned}\tag{81}$$

となる。ここで、$S_{py}^{(2)}$ を考える。

$$\begin{aligned}S_{py}^{(2)}&=\frac{1}{n}\sum_{k=1}^{n}\left(x_{kp}-\bar{x}_p\right)\left(y_k-\bar{y}\right)\\&=\frac{1}{n}\sum_{k=1}^{n}\left(x_{kp}-\bar{x}_p\right)y_k-\bar{y}\frac{1}{n}\sum_{k=1}^{n}\left(x_{kp}-\bar{x}_p\right)\\&=\frac{1}{n}\sum_{k=1}^{n}\left(x_{kp}-\bar{x}_p\right)y_k\end{aligned}\tag{82}$$

よって、

$$\begin{aligned}\hat{a}_1&=\sum_{p=1}^{2}S^{1p(2)}S_{py}^{(2)}\\&=\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{2}S^{1p(2)}\left(x_{kp}-\bar{x}_p\right)\right]y_k\\&=\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{2}S^{1p(2)}\left(x_{kp}-\bar{x}_p\right)\right]\left(a_0+a_1x_{k1}+a_2x_{k2}+e_k\right)\end{aligned}\tag{83}$$

となる。$\hat{a}_2$ も同様に以下のようになる。

$$
\begin{aligned}
\hat{a}_2 &= S^{21(2)}S_{1y}^{(2)} + S^{22(2)}S_{2y}^{(2)} \\
&= \sum_{p=1}^{2} S^{2p(2)}S_{py}^{(2)} \\
&= \frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{2} S^{2p(2)}\left(x_{kp}-\bar{x}_p\right)\right]y_k \\
&= \frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{2} S^{2p(2)}\left(x_{kp}-\bar{x}_p\right)\right]\left(a_0+a_1x_{k1}+a_2x_{k2}+e_k\right)
\end{aligned}
\tag{84}
$$

$\hat{a}_0$ は $\hat{a}_1, \hat{a}_2$ から以下のように求めることができる。

$$
\begin{aligned}
\hat{a}_0 &= \bar{y} - \left(\hat{a}_1\bar{x}_1 + \hat{a}_2\bar{x}_2\right) \\
&= \frac{1}{n}\sum_{k=1}^{n} y_k - \frac{1}{n}\sum_{k=1}^{n}\left[\bar{x}_1\sum_{p=1}^{2} S^{1p(2)}\left(x_{kp}-\bar{x}_p\right)\right]y_k - \frac{1}{n}\sum_{k=1}^{n}\left[\bar{x}_2\sum_{p=1}^{2} S^{2p(2)}\left(x_{kp}-\bar{x}_p\right)\right]y_k \\
&= \frac{1}{n}\sum_{k=1}^{n}\left\{1-\bar{x}_1\sum_{p=1}^{2} S^{1p(2)}\left(x_{kp}-\bar{x}_p\right)-\bar{x}_2\sum_{p=1}^{2} S^{2p(2)}\left(x_{kp}-\bar{x}_p\right)\right\}y_k \\
&= \frac{1}{n}\sum_{k=1}^{n}\left\{1-\sum_{j=1}^{2}\bar{x}_j\sum_{p=1}^{2} S^{jp(2)}\left(x_{kp}-\bar{x}_p\right)\right\}y_k \\
&= \frac{1}{n}\sum_{k=1}^{n}\left\{1-\sum_{j=1}^{2}\bar{x}_j\sum_{p=1}^{2} S^{jp(2)}\left(x_{kp}-\bar{x}_p\right)\right\}\left(a_0+a_1x_{k1}+a_2x_{k2}+e_k\right)
\end{aligned}
\tag{85}
$$

となる。ここで、再び e_k だけが確率変数であることに注目する。そして、期待値は以下のように評価される。

$$
\begin{aligned}
E\left[\hat{a}_1\right] &= E\left[\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{2} S^{1p(2)}\left(x_{kp}-\bar{x}_p\right)\right]\left(a_0+a_1x_{k1}+a_2x_{k2}+e_k\right)\right] \\
&= a_0\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{2} S^{1p(2)}\left(x_{kp}-\bar{x}_p\right)\right] \\
&\quad + a_1\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{2} S^{1p(2)}\left(x_{kp}-\bar{x}_p\right)\right]x_{k1} \\
&\quad + a_2\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{2} S^{1p(2)}\left(x_{kp}-\bar{x}_p\right)\right]x_{k2} \\
&\quad + E\left[\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{2} S^{1p(2)}\left(x_{kp}-\bar{x}_p\right)\right]e_k\right] \\
&= a_1\sum_{p=1}^{2} S^{1p(2)}S_{p1}^{(2)} + a_2\sum_{p=1}^{2} S^{1p(2)}S_{p2}^{(2)} \\
&= a_1
\end{aligned}
\tag{86}
$$

ここで、以下を利用している。

$$
\begin{pmatrix} S^{11(2)} & S^{12(2)} \\ S^{12(2)} & S^{22(2)} \end{pmatrix}\begin{pmatrix} S_{11}^{(2)} & S_{12}^{(2)} \\ S_{21}^{(2)} & S_{22}^{(2)} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}
\tag{87}
$$

$$
\begin{aligned}
E\left[\hat{a}_2\right] &= E\left[\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{2}S^{2p(2)}\left(x_{ip}-\bar{x}_p\right)\right]\left(a_0+a_1x_{k1}+a_2x_{k2}+e_k\right)\right] \\
&= a_0\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{2}S^{2p(2)}\left(x_{kp}-\bar{x}_p\right)\right] \\
&\quad +a_1\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{2}S^{2p(2)}\left(x_{kp}-\bar{x}_p\right)\right]x_{k1} \\
&\quad +a_2\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{2}S^{2p(2)}\left(x_{kp}-\bar{x}_p\right)\right]x_{k2} \\
&\quad +E\left[\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{2}S^{2p(2)}\left(x_{kp}-\bar{x}_p\right)\right]e_k\right] \\
&= a_1\sum_{p=1}^{2}S^{2p(2)}S_{p1}^{(2)}+a_2\sum_{p=1}^{2}S^{2p(2)}S_{p2}^{(2)} \\
&= a_2
\end{aligned}
\tag{88}
$$

$$
\begin{aligned}
E\left[\hat{a}_0\right] &= \frac{1}{n}\sum_{k=1}^{n}\left\{1-\sum_{j=1}^{2}\bar{x}_j\sum_{p=1}^{2}S^{jp(2)}\left(x_{kp}-\bar{x}_p\right)\right\}\left(a_0+a_1x_{k1}+a_2x_{k2}+e_k\right) \\
&= a_0\frac{1}{n}\sum_{k=1}^{n}\left[1-\sum_{j=1}^{2}\bar{x}_j\sum_{p=1}^{2}S^{jp(2)}\left(x_{kp}-\bar{x}_p\right)\right] \\
&\quad +a_1\frac{1}{n}\sum_{k=1}^{n}\left[1-\sum_{j=1}^{2}\bar{x}_j\sum_{p=1}^{2}S^{jp(2)}\left(x_{kp}-\bar{x}_p\right)\right]x_{k1} \\
&\quad +a_2\frac{1}{n}\sum_{k=1}^{n}\left[1-\sum_{j=1}^{2}\bar{x}_j\sum_{p=1}^{2}S^{jp(2)}\left(x_{kp}-\bar{x}_p\right)\right]x_{k2} \\
&\quad +E\left[\frac{1}{n}\sum_{k=1}^{n}\left[1-\sum_{j=1}^{2}\bar{x}_j\sum_{p=1}^{2}S^{jp(2)}\left(x_{kp}-\bar{x}_p\right)\right]e_k\right]
\end{aligned}
\tag{89}
$$

ここで、第 1 項目は

$$
\begin{aligned}
&a_0\frac{1}{n}\sum_{k=1}^{n}\left[1-\sum_{j=1}^{2}\bar{x}_j\sum_{p=1}^{2}S^{jp(2)}\left(x_{kp}-\bar{x}_p\right)\right] \\
&= a_0-a_0\sum_{k=1}^{n}\left[\bar{x}_1\sum_{p=1}^{2}S^{1p(2)}\left(x_{kp}-\bar{x}_p\right)\right]-a_0\sum_{k=1}^{n}\left[\bar{x}_2\sum_{p=1}^{2}S^{2p(2)}\left(x_{kp}-\bar{x}_p\right)\right] \\
&= a_0-a_0\left[\bar{x}_1\sum_{p=1}^{2}S^{1p(2)}\sum_{k=1}^{n}\left(x_{kp}-\bar{x}_p\right)\right]-a_0\left[\bar{x}_2\sum_{p=1}^{2}S^{2p(2)}\sum_{k=1}^{n}\left(x_{kp}-\bar{x}_p\right)\right] \\
&= a_0
\end{aligned}
\tag{90}
$$

となる。

第 2 項目は

$$
\begin{aligned}
&a_1 \frac{1}{n}\sum_{k=1}^{n}\left[1-\sum_{j=1}^{2}\bar{x}_j\sum_{p=1}^{2}S^{jp(2)}\left(x_{kp}-\bar{x}_p\right)\right]x_{k1}\\
&=a_1\bar{x}_1-a_1\frac{1}{n}\sum_{k=1}^{n}\left[\bar{x}_1\sum_{p=1}^{2}S^{1p(2)}\left(x_{kp}-\bar{x}_p\right)\right]x_{k1}-a_1\frac{1}{n}\sum_{k=1}^{n}\left[\bar{x}_2\sum_{p=1}^{2}S^{2p(2)}\left(x_{kp}-\bar{x}_p\right)\right]x_{k1}\\
&=a_1\bar{x}_1-a_1\bar{x}_1\sum_{p=1}^{2}S^{1p(2)}S_{p1}^{(2)}-a_1\bar{x}_2\sum_{p=1}^{2}S^{2p(2)}S_{p1}^{(2)}\\
&=a_1\bar{x}_1-a_1\bar{x}_1S^{11(2)}S_{11}^{(2)}\\
&=0
\end{aligned}
\tag{91}
$$

となる。

第 3 項目は

$$
\begin{aligned}
&a_2 \frac{1}{n}\sum_{k=1}^{n}\left[1-\sum_{j=1}^{2}\bar{x}_j\sum_{p=1}^{2}S^{jp(2)}\left(x_{kp}-\bar{x}_p\right)\right]x_{k2}\\
&=a_2\bar{x}_2-a_2\bar{x}_1\sum_{p=1}^{2}S^{1p(2)}\sum_{k=1}^{n}\left(x_{kp}-\bar{x}_p\right)x_{k2}-a_2\bar{x}_2\sum_{p=1}^{2}S^{2p(2)}\sum_{k=1}^{n}\left(x_{kp}-\bar{x}_p\right)x_{k2}\\
&=a_2\bar{x}_2-a_2\bar{x}_1\sum_{p=1}^{2}S^{1p(2)}S_{p2}^{(2)}-a_2\bar{x}_2\sum_{p=1}^{2}S^{2p(2)}S_{p2}^{(2)}\\
&=a_2\bar{x}_2-a_2\bar{x}_2S^{22(2)}S_{22}^{(2)}\\
&=a_2\bar{x}_2-a_2\bar{x}_2\\
&=0
\end{aligned}
\tag{92}
$$

となる。

第 4 項目は

$$
E\left[\frac{1}{n}\sum_{k=1}^{n}\left[1-\sum_{j=1}^{2}\bar{x}_j\sum_{p=1}^{2}S^{jp(2)}\left(x_{kp}-\bar{x}_p\right)\right]e_k\right]=0 \tag{93}
$$

となる。したがって、以下を得る。

$$
E\left[\hat{a}_0\right]=a_0 \tag{94}
$$

次に分散を評価する。それは以下のようになる。

$$
\begin{aligned}
V\left[\hat{a}_1\right] &= \sum_{k=1}^{n}\left[\frac{1}{n}\sum_{p=1}^{2} S^{1p(2)}\left(x_{kp}-\bar{x}_p\right)\right]^2 V\left[e_k\right] \\
&= \frac{1}{n}\frac{1}{n}\sum_{k=1}^{n}\left[S^{11(2)}\left(x_{k1}-\bar{x}_1\right)+S^{12(2)}\left(x_{k2}-\bar{x}_2\right)\right]^2 V\left[e_k\right] \\
&= \frac{1}{n}\frac{1}{n}\sum_{k=1}^{n}\left[\left(S^{11(2)}\right)^2\left(x_{k1}-\bar{x}_1\right)^2+2S^{12(2)}S^{11(2)}\left(x_{k1}-\bar{x}_1\right)\left(x_{k2}-\bar{x}_2\right)+\left(S^{12(2)}\right)^2\left(x_{k2}-\bar{x}_2\right)^2\right]s_e^{(2)} \\
&= \frac{1}{n}\left[\left(S^{11(2)}\right)^2 S_{11}^{(2)}+2S^{12(2)}S^{11(2)}S_{12}^{(2)}+\left(S^{12(2)}\right)^2 S_{22}^{(2)}\right]s_e^{(2)} \\
&= \frac{1}{n}\left[S^{11(2)}\left[S^{11(2)}S_{11}^{(2)}+S^{12(2)}S_{21}^{(2)}\right]+S^{12(2)}\left[S^{11(2)}S_{12}^{(2)}+S^{12(2)}S_{22}^{(2)}\right]\right]s_e^{(2)} \\
&= \frac{S^{11(2)}}{n}s_e^{(2)}
\end{aligned}
\tag{95}
$$

$$
\begin{aligned}
V\left[\hat{a}_2\right] &= \sum_{k=1}^{n}\left[\frac{1}{n}\sum_{p=1}^{2} S^{2p(2)}\left(x_{kp}-\bar{x}_p\right)\right]^2 V\left[e\right] \\
&= \frac{1}{n}\frac{1}{n}\sum_{k=1}^{n}\left[S^{21(2)}\left(x_{k1}-\bar{x}_1\right)+S^{22(2)}\left(x_{k2}-\bar{x}_2\right)\right]^2 s_e^{(2)} \\
&= \frac{1}{n}\frac{1}{n}\sum_{k=1}^{n}\left[\left(S^{21(2)}\right)^2\left(x_{k2}-\bar{x}_2\right)^2+2S^{21(2)}S^{22(2)}\left(x_{k1}-\bar{x}_1\right)\left(x_{k2}-\bar{x}_2\right)+\left(S^{22(2)}\right)^2\left(x_{k2}-\bar{x}_2\right)^2\right]^2 s_e^{(2)} \\
&= \frac{1}{n}\left[\left(S^{21(2)}\right)^2 S_{22}^{(2)}+2S^{21(2)}S^{22(2)}S_{12}^{(2)}+\left(S^{22(2)}\right)^2 S_{22}^{(2)}\right]s_e^{(2)} \\
&= \frac{1}{n}\left[S^{21(2)}\left[S^{21(2)}S_{22}^{(2)}+S^{22(2)}S_{12}^{(2)}\right]+S^{22(2)}\left[S^{21(2)}S_{12}^{(2)}+S^{22(2)}S_{22}^{(2)}\right]\right]s_e^{(2)} \\
&= \frac{S^{22(2)}}{n}s_e^{(2)}
\end{aligned}
\tag{96}
$$

$$
\begin{aligned}
V\left[\hat{a}_0\right] &= V\left[\frac{1}{n}\sum_{k=1}^{n}\left\{1-\sum_{j=1}^{2}\bar{x}_j\sum_{p=1}^{2}S^{jp(2)}\left(x_{kp}-\bar{x}_p\right)\right\}\left(a_0+a_1x_{k1}+a_2x_{k2}+e_i\right)\right] \\
&= \left[\sum_{k=1}^{n}\left[\frac{1}{n}\left\{1-\sum_{j=1}^{2}\bar{x}_j\sum_{p=1}^{2}S^{jp(2)}\left(x_{kp}-\bar{x}_p\right)\right\}\right]^2\right]V\left[e\right] \\
&= \frac{1}{n}\frac{1}{n}\left[\sum_{k=1}^{n}\left[\left\{1-\sum_{j=1}^{2}\bar{x}_j\left[S^{j1(2)}\left(x_{k1}-\bar{x}_1\right)+S^{j2(2)}\left(x_{k2}-\bar{x}_2\right)\right]\right\}^2\right]\right]V\left[e\right] \\
&= \frac{1}{n}\frac{1}{n}\left[\sum_{k=1}^{n}\left[1-\left\{\bar{x}_1\left[S^{11(2)}\left(x_{k1}-\bar{x}_1\right)+S^{12(2)}\left(x_{k2}-\bar{x}_2\right)\right]+\bar{x}_2\left[S^{21(2)}\left(x_{k1}-\bar{x}_1\right)+S^{22(2)}\left(x_{k2}-\bar{x}_2\right)\right]\right\}^2\right]\right]s_e^{(2)} \\
&= \frac{1}{n}\frac{1}{n}\left[\sum_{k=1}^{n}\left[\begin{array}{l}1+\left\{\bar{x}_1\left[S^{11(2)}\left(x_{k1}-\bar{x}_1\right)+S^{12(2)}\left(x_{k2}-\bar{x}_2\right)\right]+\bar{x}_2\left[S^{21(2)}\left(x_{k1}-\bar{x}_1\right)+S^{22(2)}\left(x_{k2}-\bar{x}_2\right)\right]\right\}^2 \\ -2\left\{\bar{x}_1\left[S^{11(2)}\left(x_{k1}-\bar{x}_1\right)+S^{12(2)}\left(x_{k2}-\bar{x}_2\right)\right]+\bar{x}_2\left[S^{21(2)}\left(x_{k1}-\bar{x}_1\right)+S^{22(2)}\left(x_{k2}-\bar{x}_2\right)\right]\right\}\end{array}\right]\right]s_e^{(2)} \\
&= \frac{1}{n}\frac{1}{n}\left[\sum_{k=1}^{n}\left[1+\left\{\begin{array}{l}\bar{x}_1^{\,2}\left[S^{11(2)}\left(x_{k1}-\bar{x}_1\right)+S^{12(2)}\left(x_{k2}-\bar{x}_2\right)\right]^2+\bar{x}_2^{\,2}\left[S^{21(2)}\left(x_{k1}-\bar{x}_1\right)+S^{22(2)}\left(x_{k2}-\bar{x}_2\right)\right]^2 \\ +2\bar{x}_1\bar{x}_2\left[S^{11(2)}\left(x_{k1}-\bar{x}_1\right)+S^{12(2)}\left(x_{k2}-\bar{x}_2\right)\right]\left[S^{21(2)}\left(x_{k1}-\bar{x}_1\right)+S^{22(2)}\left(x_{k2}-\bar{x}_2\right)\right]\end{array}\right\}\right]\right]s_e^{(2)} \\
&= \frac{1}{n}+\frac{1}{n}\frac{1}{n}\left[\sum_{k=1}^{n}\left[\begin{array}{l}\bar{x}_1^{\,2}\left[\left(S^{11(2)}\right)^2\left(x_{k1}-\bar{x}_1\right)^2+\left(S^{12(2)}\right)^2\left(x_{k2}-\bar{x}_2\right)^2+2S^{11(2)}S^{12(2)}\left(x_{k1}-\bar{x}_1\right)\left(x_{k2}-\bar{x}_2\right)\right] \\ +\bar{x}_2^{\,2}\left[\left(S^{21(2)}\right)^2\left(x_{k1}-\bar{x}_1\right)^2+\left(S^{22(2)}\right)^2\left(x_{k2}-\bar{x}_2\right)^2+2S^{21(2)}S^{22(2)}\left(x_{k1}-\bar{x}_1\right)\left(x_{k2}-\bar{x}_2\right)\right] \\ +2\bar{x}_1\bar{x}_2\left[\begin{array}{l}S^{11(2)}S^{21(2)}\left(x_{k1}-\bar{x}_1\right)^2+S^{11(2)}S^{22(2)}\left(x_{k1}-\bar{x}_1\right)\left(x_{k2}-\bar{x}_2\right) \\ +S^{12(2)}S^{21(2)}\left(x_{k2}-\bar{x}_2\right)\left(x_{k1}-\bar{x}_1\right)+S^{12(2)}S^{22(2)}\left(x_{k2}-\bar{x}_2\right)\left(x_{k2}-\bar{x}_2\right)\end{array}\right]\end{array}\right]\right]s_e^{(2)} \\
&= \left\{\frac{1}{n}+\frac{1}{n}\left[\begin{array}{l}\bar{x}_1^{\,2}\left[\left(S^{11(2)}\right)^2S_{11}^{(2)}+\left(S^{12(2)}\right)^2S_{22}^{(2)}+2S^{11(2)}S^{12(2)}S_{12}^{(2)}\right] \\ +\bar{x}_2^{\,2}\left[\left(S^{21(2)}\right)^2S_{11}^{(2)}+\left(S^{22(2)}\right)^2S_{22}^{(2)2}+2S^{21(2)}S^{22(2)}S_{12}^{(2)}\right] \\ +2\bar{x}_1\bar{x}_2\left[\begin{array}{l}S^{11(2)}\left[S^{21(2)}S_{11}^{(2)}+S^{22(2)}S_{12}^{(2)}\right] \\ +S^{12(2)}\left[S^{21(2)}S_{21}^{(2)}+S^{22(2)}S_{22}^{(2)}\right]\end{array}\right]\end{array}\right]\right\}s_e^{(2)} \\
&= \left\{\frac{1}{n}+\frac{1}{n}\left[\bar{x}_1^{\,2}\left[S^{11(2)}\right]+\bar{x}_2^{\,2}\left[S^{22(2)}\right]+2\bar{x}_1\bar{x}_2S^{12(2)}\right]\right\}s_e^{(2)} \\
&= \left\{\frac{1}{n}+\frac{1}{n}\left[\bar{x}_1^{\,2}\left[S^{11(2)}\right]+\bar{x}_2^{\,2}\left[S^{22(2)}\right]+\bar{x}_1\bar{x}_2\left[S^{12(2)}+S^{21(2)}\right]\right]\right\}s_e^{(2)} \\
&= \frac{1}{n}\left[1+\sum_{j=1}^{2}\sum_{l=1}^{2}\bar{x}_j\bar{x}_lS^{jl(2)}\right]s_e^{(2)} \\
&= \frac{1}{n}\left\{1+\begin{pmatrix}\bar{x}_1 & \bar{x}_2\end{pmatrix}\begin{pmatrix}S^{11(2)} & S^{12(2)} \\ S^{21(2)} & S^{22(2)}\end{pmatrix}\begin{pmatrix}\bar{x}_1 \\ \bar{x}_2\end{pmatrix}\right\}s_e^{(2)}
\end{aligned}
$$

(97)

ここで、以下を利用している。

$$S^{12(2)} = S^{21(2)} \tag{98}$$

$\hat{a}_1$ と $\hat{a}_2$ 間の共分散の期待値は以下のように求まる

$$\begin{aligned}
Cov[\hat{a}_1, \hat{a}_2] &= \frac{1}{n}\frac{1}{n}\left\{\sum_{k=1}^{n}\left[\sum_{p=1}^{2} S^{1p(2)}\left(x_{kp} - \bar{x}_p\right)\right]\left[\sum_{p'=1}^{2} S^{2p'(2)}\left(x_{kp'} - \bar{x}_{p'}\right)\right]\right\} V[e] \\
&= \frac{1}{n}\frac{1}{n}\left\{\sum_{k=1}^{n}\left[S^{11(2)}\left(x_{k1} - \bar{x}_1\right) + S^{12(2)}\left(x_{k2} - \bar{x}_2\right)\right]\left[S^{21(2)}\left(x_{k1} - \bar{x}_1\right) + S^{22(2)}\left(x_{k2} - \bar{x}_2\right)\right]\right\} s_e^{(2)} \\
&= \frac{1}{n}\frac{1}{n}\sum_{k=1}^{n}\left[\begin{array}{l} S^{11(2)}S^{21(2)}\left(x_{k1} - \bar{x}_1\right)^2 + S^{11(2)}S^{22(2)}\left(x_{k1} - \bar{x}_1\right)\left(x_{k2} - \bar{x}_2\right) \\ +S^{12(2)}S^{21(2)}\left(x_{k2} - \bar{x}_2\right)\left(x_{k1} - \bar{x}_1\right) + S^{12(2)}S^{22(2)}\left(x_{k2} - \bar{x}_2\right)^2 \end{array}\right] s_e^{(2)} \\
&= \frac{1}{n}\left[\begin{array}{l} S^{11(2)}\left[S^{21(2)}S_{11}^{(2)} + S^{22(2)}S_{12}^{(2)}\right] \\ +S^{12(2)}\left[S^{21(2)}S_{21}^{(2)} + S^{22(2)}S_{22}^{(2)}\right] \end{array}\right] s_e^{(2)} \\
&= \frac{1}{n} S^{12(2)} s_e^{(2)}
\end{aligned} \tag{99}$$

同様に

$$Cov[\hat{a}_2, \hat{a}_1] = \frac{1}{n} S^{21(2)} s_e^{(2)} \tag{100}$$

となる。一般に $\hat{a}_i$ と $\hat{a}_j$ $(i, j \neq 0)$ 間の共分散は

$$Cov[\hat{a}_j, \hat{a}_l] = \frac{1}{n} S^{jl(2)} s_e^{(2)} \tag{101}$$

となる。

$\hat{a}_0$ と $\hat{a}_1$ 間の共分散の期待値は

$$
\begin{aligned}
Cov\left[\hat{a}_0,\hat{a}_1\right] &= \frac{1}{n}\frac{1}{n}\sum_{k=1}^{n}\left\{1-\sum_{j=1}^{2}\bar{x}_j\sum_{p=1}^{2}S^{jp(2)}\left(x_{kp}-\bar{x}_p\right)\right\}\left[\sum_{p=1}^{2}S^{1p(2)}\left(x_{kp}-\bar{x}_p\right)\right]V\left[e\right] \\
&= \frac{1}{n}\frac{1}{n}\sum_{k=1}^{n}\left\{1-\sum_{j=1}^{2}\bar{x}_j\left[S^{j1(2)}\left(x_{k1}-\bar{x}_1\right)+S^{j2(2)}\left(x_{k2}-\bar{x}_2\right)\right]\right\}\left[S^{11(2)}\left(x_{k1}-\bar{x}_1\right)+S^{12(2)}\left(x_{k2}-\bar{x}_2\right)\right]s_e^{(2)} \\
&= \frac{1}{n}\frac{1}{n}\sum_{k=1}^{n}\begin{Bmatrix}1-\bar{x}_1\left[S^{11(2)}\left(x_{k1}-\bar{x}_1\right)+S^{12(2)}\left(x_{k2}-\bar{x}_2\right)\right] \\ -\bar{x}_2\left[S^{21(2)}\left(x_{k1}-\bar{x}_1\right)+S^{22(2)}\left(x_{k2}-\bar{x}_2\right)\right]\end{Bmatrix}\left[S^{11(2)}\left(x_{k1}-\bar{x}_1\right)+S^{12(2)}\left(x_{k2}-\bar{x}_2\right)\right]s_e^{(2)} \\
&= -\frac{1}{n}\frac{1}{n}\sum_{k=1}^{n}\begin{Bmatrix}\bar{x}_1\left[S^{11(2)}\left(x_{k1}-\bar{x}_1\right)+S^{12(2)}\left(x_{k2}-\bar{x}_2\right)\right] \\ +\bar{x}_2\left[S^{21(2)}\left(x_{k1}-\bar{x}_1\right)+S^{22(2)}\left(x_{k2}-\bar{x}_2\right)\right]\end{Bmatrix}\left[S^{11(2)}\left(x_{k1}-\bar{x}_1\right)+S^{12(2)}\left(x_{k2}-\bar{x}_2\right)\right]s_e^{(2)} \\
&= -\frac{1}{n}\frac{1}{n}\sum_{k=1}^{n}\begin{Bmatrix}\bar{x}_1\left[S^{11(2)}S^{11(2)}\left(x_{k1}-\bar{x}_1\right)^2+S^{12(2)}S^{11(2)}\left(x_{k2}-\bar{x}_2\right)\left(x_{k1}-\bar{x}_1\right)\right] \\ +\bar{x}_2\left[S^{21(2)}S^{11(2)}\left(x_{k1}-\bar{x}_1\right)^2+S^{22(2)}S^{11(2)}\left(x_{ik}-\bar{x}_2\right)\left(x_{k1}-\bar{x}_1\right)\right] \\ \bar{x}_1\left[S^{11(2)}S^{12(2)}\left(x_{k1}-\bar{x}_1\right)\left(x_{k2}-\bar{x}_2\right)+S^{12(2)}S^{12(2)}\left(x_{k2}-\bar{x}_2\right)^2\right] \\ +\bar{x}_2\left[S^{21(2)}S^{12(2)}\left(x_{k1}-\bar{x}_1\right)\left(x_{k2}-\bar{x}_2\right)+S^{22(2)}S^{12(2)}\left(x_{k2}-\bar{x}_2\right)^2\right]\end{Bmatrix}s_e^{(2)} \\
&= -\frac{1}{n}\begin{Bmatrix}\bar{x}_1\left[S^{11(2)}S^{11(2)}S_{11}^{(2)}+S^{12(2)}S^{11(2)}S_{21}^{(2)}\right] \\ +\bar{x}_2\left[S^{21(2)}S^{11(2)}S_{11}^{(2)}+S^{22(2)}S^{11(2)}S_{21}^{(2)}\right] \\ \bar{x}_1\left[S^{11(2)}S^{12(2)}S_{12}^{(2)}+S^{12(2)}S^{12(2)}S_{22}^{(2)}\right] \\ +\bar{x}_2\left[S^{21(2)}S^{12(2)}S_{12}^{(2)}+S^{22(2)}S^{12(2)}S_{22}^{(2)}\right]\end{Bmatrix}s_e^{(2)} \\
&= -\frac{1}{n}\left\{\bar{x}_1S^{11(2)}+\bar{x}_2S^{12(2)}\right\}s_e^{(2)} \\
&= -\frac{1}{n}\begin{pmatrix}S^{11(2)} & S^{12(2)}\end{pmatrix}\begin{pmatrix}\bar{x}_1\\ \bar{x}_2\end{pmatrix}
\end{aligned}
\tag{102}
$$

となる。

同様に $\hat{a}_0$ と $\hat{a}_2$ 間の共分散の期待値は

$$
\begin{aligned}
Cov\left[\hat{a}_0,\hat{a}_2\right] &= -\frac{1}{n}\left\{\bar{x}_1S^{21(2)}+\bar{x}_2S^{22(2)}\right\}s_e^{(2)} \\
&= -\frac{1}{n}\begin{pmatrix}S^{21(2)} & S^{22(2)}\end{pmatrix}\begin{pmatrix}\bar{x}_1\\ \bar{x}_2\end{pmatrix}s_e^{(2)}
\end{aligned}
\tag{103}
$$

となる。結局 $Cov\left[\hat{a}_0,\hat{a}_j\right]$ は、一般に

$$
Cov\left[\hat{a}_0,\hat{a}_j\right] = -\frac{1}{n}\sum_{l=1}^{2}\bar{x}_lS^{jl(2)}s_e^{(2)} \tag{104}
$$

となる。

$\hat{a}_1$ と $\hat{a}_0$ は正規分布に従うとする。したがって、その規格化値

$$t = \frac{\hat{a}_i - a_i}{\sqrt{V[a_i]}} \quad for\ i = 0,1,2 \tag{105}$$

は自由度 ϕ_e の t 分布に従う。ただし、自由度は

$$\phi_e = n - 3 \tag{106}$$

である。したがって、係数の推定範囲は

$$\hat{a}_i - t_p(\phi_e;P)V[\hat{a}_i] \le a_i \le \hat{a}_i + t_p(\phi_e;P)V[\hat{a}_i] \tag{107}$$

となる。

一方、その変数が意味をなすかは、その係数が 0 と判定できるかで判断できるであろう。すなわち、変数 x_i が意味があるかは、対応する係数が

$$t_i = \frac{\hat{a}_i}{\sqrt{\frac{S^{ii(2)}}{n} s_e^{(2)}}} \tag{108}$$

が、0 でないとき、すなわち

$$t_i > t_P\left(n-(p+1)\right) = t_P\left(n-(2+1)\right) \tag{109}$$

であるとき意味がある、と判断できる。

4.2.8. 回帰値および目的変数値の推定範囲

ここでは、回帰直線および、データの誤差範囲を検討する。説明変数の値 x_1, x_2 に対応する回帰値を Y とする。

実際に説明変数と目的変数の値を得た場合、それと回帰値と説明変数の値の範囲を比較することになる。実際の値と範囲の値の比較は以下の意味を持つ。

実際の値と回帰値の範囲の比較は、その平均との差に相当する。したがって、その上限よりも大きければ、平均よりも実際の値は大きく、その下限よりも小さければ、平均よりも実際の値は小さいと判断される。

実際の値と目的変数の範囲の比較は、そのデータが妥当範囲であるかを判断できる。目的変数の上限よりも大きければ、それは相場よりも大きいことを意味し、目的変数の下限よりも小さければ、それは相場よりも小さいことを意味する。

我々は n セットのデータ (x_{k1}, x_{k2}, y_k) を得たとする。対応する回帰値は以下で与えられる。

$$Y = \hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2 \tag{110}$$

この期待値は

$$\begin{aligned} E[Y] &= E[\hat{a}_0] + E[\hat{a}_1] x_1 + E[\hat{a}_2] x_2 \\ &= a_0 + a_1 x_1 + a_2 x_2 \end{aligned} \tag{111}$$

となる。

Y の分散は

$$\begin{aligned} V[Y] &= V[\hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2] \\ &= V[\hat{a}_0] + 2\sum_{j=1}^{2} x_j Cov[\hat{a}_0, \hat{a}_j] + \sum_{j=1}^{2}\sum_{l=1}^{2} x_j x_l Cov[\hat{a}_j, \hat{a}_l] \\ &= \frac{1}{n}\left[1 + \sum_{j=1}^{2}\sum_{l=1}^{2} \bar{x}_j \bar{x}_l S^{jl(2)}\right] s_e^{(2)} - \frac{2}{n}\left[\sum_{j=1}^{2} x_j \sum_{l=1}^{2} \bar{x}_l S^{jl(2)}\right] s_e^{(2)} + \frac{1}{n}\left[\sum_{j=1}^{2}\sum_{l=1}^{2} x_j x_l S^{jl(2)}\right] s_e^{(2)} \\ &= \frac{1}{n}\left[1 + \sum_{j=1}^{2}\sum_{l=1}^{2} \left(\bar{x}_j \bar{x}_l - 2x_j \bar{x}_l + x_j x_l\right) S^{jl(2)}\right] s_e^{(2)} \end{aligned} \tag{112}$$

となる。ここで、以下を考える。

$$\begin{aligned} &\sum_{j=1}^{2}\sum_{l=1}^{2} 2x_j \bar{x}_l S^{jl(2)} \\ &= \sum_{j=1}^{2}\sum_{l=1}^{2} x_j \bar{x}_l S^{jl(2)} + \sum_{j=1}^{2}\sum_{l=1}^{2} x_j \bar{x}_l S^{jl(2)} \end{aligned} \tag{113}$$

二番目の項は以下のように変形される。

$$\begin{aligned} \sum_{j=1}^{2}\sum_{l=1}^{2} x_j \bar{x}_l S^{jl(2)} &= \sum_{j=1}^{2}\sum_{l=1}^{2} x_l \bar{x}_j S^{lj(2)} \\ &= \sum_{j=1}^{2}\sum_{l=1}^{2} x_l \bar{x}_j S^{jl(2)} \end{aligned} \tag{114}$$

ここで、

$$S^{lj(2)} = S^{jl(2)} \tag{115}$$

を利用している。したがって、我々は以下を得る。

$$\sum_{j=1}^{2}\sum_{l=1}^{2} 2x_j \bar{x}_l S^{jl(2)} = \sum_{j=1}^{2}\sum_{l=1}^{2} \left(x_j \bar{x}_l + x_l \bar{x}_j\right) S^{jl(2)} \tag{116}$$

この Eq. (116)を Eq. (112)に代入して以下を得る。

$$
\begin{aligned}
V[Y] &= \frac{1}{n}\left[1+\sum_{j=1}^{2}\sum_{l=1}^{2}\left(\bar{x}_j\bar{x}_l - 2x_j\bar{x}_l + x_jx_l\right)S^{jl(2)}\right]s_e^{(2)} \\
&= \frac{1}{n}\left[1+\sum_{j=1}^{2}\sum_{l=1}^{2}\left(\bar{x}_j\bar{x}_l - x_j\bar{x}_l - x_l\bar{x}_j + x_jx_l\right)S^{jl(2)}\right]s_e^{(2)} \\
&= \frac{1}{n}\left[1+\sum_{j=1}^{2}\sum_{l=1}^{2}\left(x_j-\bar{x}_j\right)\left(x_l-\bar{x}_l\right)S^{jl(2)}\right]s_e^{(2)} \\
&= \frac{1}{n}\left[1+\begin{pmatrix}x_1-\bar{x}_1 & x_2-\bar{x}_2\end{pmatrix}\begin{pmatrix}S^{11(2)} & S^{12(2)}\\ S^{21(2)} & S^{22(2)}\end{pmatrix}\begin{pmatrix}x_1-\bar{x}_1\\ x_2-\bar{x}_2\end{pmatrix}\right]s_e^{(2)} \\
&= \frac{1}{n}\left(1+D^2\right)s_e^{(2)}
\end{aligned}
\tag{117}
$$

ここで、

$$
D^2 = \begin{pmatrix}x_1-\bar{x}_1 & x_2-\bar{x}_2\end{pmatrix}\begin{pmatrix}S^{11(2)} & S^{12(2)}\\ S^{21(2)} & S^{22(2)}\end{pmatrix}\begin{pmatrix}x_1-\bar{x}_1\\ x_2-\bar{x}_2\end{pmatrix} \tag{118}
$$

である。したがって、変数は以下のように規格化できる。

$$
\begin{aligned}
t &= \frac{Y-E[Y]}{V[Y]} \\
&= \frac{\left(\hat{a}_0+\hat{a}_1x_1+\hat{a}_2x_2\right)-\left(a_0+a_1x_1+a_2x_2\right)}{\sqrt{\left(\frac{1}{n}+\frac{D^2}{n}\right)s_e^{(2)}}}
\end{aligned}
\tag{119}
$$

そして、それは自由度ϕ_eの 分布に従う。したがって、回帰値の推定範囲は

$$
\hat{a}_0+\hat{a}_1x_1+\hat{a}_2x_2 - t_p\left(\phi_e;P\right)\sqrt{\left(\frac{1}{n}+\frac{D^2}{n}\right)s_e^{(2)}} \le Y \le \hat{a}_0+\hat{a}_1x_1+\hat{a}_2x_2 + t_p\left(\phi_e;P\right)\sqrt{\left(\frac{1}{n}+\frac{D^2}{n}\right)s_e^{(2)}}
$$

(120)

となる。

目的変数yは以下のように表現される。

$$
y = \hat{a}_0+\hat{a}_1x_1+\hat{a}_2x_2+e \tag{121}
$$

したがって、その推定範囲は以下のようになる。

$$
\hat{a}_0+\hat{a}_1x_1+\hat{a}_2x_2 - t\left(\phi_e,P\right)\sqrt{\left(1+\frac{1}{n}+\frac{D^2}{n}\right)s_e^{(2)}} \le y \le \hat{a}_0+\hat{a}_1x_1+\hat{a}_2x_2 + t\left(\phi_e,P\right)\sqrt{\left(1+\frac{1}{n}+\frac{D^2}{n}\right)s_e^{(2)}}
$$

(122)

4.2.9. 残差誤差

決定係数、自由度調整済決定係数、F 値、重相関係数はデータ全体を評価するものであった。ここで、個別のデータがいいのかを判断する方法を示す。

残差誤差は

$$e_k = y_k - Y_k \tag{123}$$

で評価される。これは規格化して

$$z_{ek} = \frac{e_k}{\sqrt{s_e^{(2)}}} \tag{124}$$

となる。ここで、$s_e^{(2)}$ は

$$s_e^{(2)} = \frac{n}{\phi_e} S_e^{(2)} \tag{125}$$

で与えられる。この z_{ek} は標準正規分布 $N\left[0,1^2\right]$ に従う。したがって、これから、この誤差がある有意値 z_P より大きいかどうかで、データを判断できる。このデータ評価では、どの位置のデータなのかの情報を利用していない。つまり、データ位置に関わりなくその誤差を評価している。

しかしながら、一般にデータはその平均値のまわりに密集している。したがって、平均値の近くのデータはそれぞれがずれたとしても、回帰直線を求める際に大きな影響は及ぼさない。しかし、平均値から遠ざかると、データ点の数は小さくなり、その誤差の回帰直線に与える影響は大きくなる。このデータの位置による影響を考慮したい。それは、次節で示す梃子比で表現する。

4.2.10. 梃子比

位置によるデータの誤差を考慮するため、別な視点から評価する。

k 番目の回帰値は以下のように与えられる。

$$\begin{aligned} Y_k &= \hat{a}_0 + \hat{a}_1 x_{k1} + \hat{a}_2 x_{k2} \\ &= \bar{y} - \hat{a}_1 \left(x_{k1} - \bar{x}_1 \right) + \hat{a}_2 \left(x_{k2} - \bar{x}_2 \right) \\ &= \bar{y} - \left(x_{k1} - \bar{x}_1, x_{k2} - \bar{x}_2 \right) \begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \end{pmatrix} \end{aligned} \tag{126}$$

右辺の項はそれぞれ以下のように展開される。

$$\bar{y} = \frac{1}{n} \sum_{k=1}^{n} y_k \tag{127}$$

$$
\begin{aligned}
\left(x_{k1}-\bar{x}_1, x_{k2}-\bar{x}_2\right)\begin{pmatrix}\hat{a}_1\\ \hat{a}_2\end{pmatrix} &= \left(x_{k1}-\bar{x}_1, x_{k2}-\bar{x}_2\right)\begin{pmatrix}S^{11(2)} & S^{12(2)}\\ S^{21(2)} & S^{22(2)}\end{pmatrix}\begin{pmatrix}S_{1y}^{(2)}\\ S_{2y}^{(2)}\end{pmatrix}\\
&= \left(x_{k1}-\bar{x}_1, x_{k2}-\bar{x}_2\right)\begin{pmatrix}S^{11(2)} & S^{12(2)}\\ S^{21(2)} & S^{22(2)}\end{pmatrix}\begin{pmatrix}\frac{1}{n}\sum_{k=1}^{n}\left(x_{k1}-\bar{x}_1\right)\left(y_k-\bar{y}\right)\\ \frac{1}{n}\sum_{k=1}^{n}\left(x_{k2}-\bar{x}_2\right)\left(y_k-\bar{y}\right)\end{pmatrix}\\
&= \left(x_{k1}-\bar{x}_1, x_{k2}-\bar{x}_2\right)\begin{pmatrix}S^{11(2)} & S^{12(2)}\\ S^{21(2)} & S^{22(2)}\end{pmatrix}\begin{pmatrix}\frac{1}{n}\sum_{k=1}^{n}\left(x_{k1}-\bar{x}_1\right)y_k\\ \frac{1}{n}\sum_{k=1}^{n}\left(x_{k2}-\bar{x}_2\right)y_k\end{pmatrix}
\end{aligned}
\tag{128}
$$

したがって、回帰値は

$$
Y_k = h_{k1}y_1 + h_{k2}y_2 + \cdots + h_{kk}y_k + \cdots + h_{kn}y_n \tag{129}
$$

と表現される。ここで、

$$
h_{kl} = \frac{1}{n} + \frac{1}{n}\left(x_{k1}-\bar{x}_1, x_{k2}-\bar{x}_2\right)\begin{pmatrix}S^{11(2)} & S^{12(2)}\\ S^{21(2)} & S^{22(2)}\end{pmatrix}\begin{pmatrix}x_{l1}-\bar{x}_1\\ x_{l2}-\bar{x}_2\end{pmatrix} \tag{130}
$$

である。ここで、k 番目の係数すなわち $h_{k,l=k} = h_{kk}$ に注目する。それは

$$
h_{kk} = \frac{1}{n} + \frac{D_k^2}{n} \tag{131}
$$

で与えられる。ここで、

$$
\begin{aligned}
D_k^2 &= \begin{pmatrix}x_{k1}-\bar{x}_1 & x_{k2}-\bar{x}_2\end{pmatrix}\begin{pmatrix}S^{11(2)} & S^{12(2)}\\ S^{12(2)} & S^{22(2)}\end{pmatrix}\begin{pmatrix}x_{k1}-\bar{x}_1\\ x_{k2}-\bar{x}_2\end{pmatrix}\\
&= \begin{pmatrix}x_{k1}-\bar{x}_1 & x_{k1}-\bar{x}\end{pmatrix}\begin{pmatrix}\left(x_{k1}-\bar{x}_1\right)S^{11(2)} + \left(x_{k2}-\bar{x}_2\right)S^{12(2)}\\ \left(x_{k1}-\bar{x}_1\right)S^{12(2)} + \left(x_{k2}-\bar{x}_2\right)S^{22(2)}\end{pmatrix}\\
&= \left(x_{k1}-\bar{x}_1\right)^2 S^{11(2)} + 2\left(x_{k1}-\bar{x}_1\right)\left(x_{k2}-\bar{x}_2\right)S^{12(2)} + \left(x_{k2}-\bar{x}_2\right)^2 S^{22(2)}
\end{aligned}
\tag{132}
$$

である。$D_k{}^2$ はマハラノビスの距離の平方と呼ばれる。

h_{kk} は k 番目の回帰値と k 番目の目的変数の関係を表している。具体的には h_{kk} は y_k のみが 1 変化したとき、回帰値の変動する量である。回帰値は、すべてのデータから決まって欲しい。したがって、h_{kk} は小さいほうが、回帰はうまくいっていると考えることができる。

この梃子比 h_{kk} は梃子比は x_k の平均からの距離のみで決まっており、目的変数とは関連していない。平均からの距離、つまり偏差の絶対値が大きくなるほど大きくなる。したがって、平均から遠いデータほど梃子比が大きくなる。すなわち、平均から大きくずれた点まで回帰をしてはならないことがわかる。これは、位置に関する重みと考えることができる。

梃子比に関しては以下の解析が成り立つ。

$$
\begin{aligned}
\sum_{k=1}^{n} h_{kk} &= \sum_{k=1}^{n} \frac{1}{n} + \sum_{k=1}^{n} \frac{D_k^2}{n} \\
&= 1 + \frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{2} \sum_{j=1}^{2} \left(x_{ki} - \bar{x}_i\right)\left(x_{kj} - \bar{x}_j\right) S^{ij(2)} \\
&= 1 + \sum_{i=1}^{2} \sum_{i=1}^{2} \left[\frac{\sum_{k=1}^{n} \left(x_{ki} - \bar{x}_i\right)\left(x_{kj} - \bar{x}_j\right)}{n} \right] S^{ij(2)} \\
&= 1 + \sum_{i=1}^{2} \sum_{i=1}^{2} S_{ij}^{(2)} S^{ij(2)} \\
&= 1 + 2
\end{aligned}
\tag{133}
$$

したがって、梃子比の平均は

$$
\mu_{hkk} = \frac{3}{n} \tag{134}
$$

となる。梃子比 h_{kk} の臨界値として

$$
h_{kk} \le \kappa \times \mu_{hkk} \tag{135}
$$

が採用される場合がある。$\kappa = 2$ が良く用いられる。梃子比が $\kappa\mu_{h_{kk}}$ よりも大きい場合は外れ値として扱われる場合がある。

もしも、変数 x_1 と x_2 が独立である場合、逆行列は以下のように単純になる。

$$
\begin{aligned}
\begin{pmatrix} S^{11(2)} & S^{12(2)} \\ S^{12(2)} & S^{22(2)} \end{pmatrix} &= \begin{pmatrix} S_{11}^{(2)} & S_{12}^{(2)} \\ S_{12}^{(2)} & S_{22}^{(2)} \end{pmatrix}^{-1} \\
&= \begin{pmatrix} \dfrac{1}{S_{11}^{(2)}} & 0 \\ 0 & \dfrac{1}{S_{22}^{(2)}} \end{pmatrix}
\end{aligned}
\tag{136}
$$

そして、マハラノビスの距離の平方は

$$
D_k^2 = \frac{\left(x_{k1} - \bar{x}_1\right)^2}{S_{11}^{(2)}} + \frac{\left(x_{k2} - \bar{x}_2\right)^2}{S_{22}^{(2)}} \tag{137}
$$

これは、梃子比では単純に二項を加えていけばいいことになる。

$$
h_{kk} = \frac{1}{n} + \frac{\left(x_{k1} - \bar{x}_1\right)^2}{nS_{11}^{(2)}} + \frac{\left(x_{k2} - \bar{x}_2\right)^2}{nS_{22}^{(2)}} \tag{138}
$$

4.2.11. 誤差分散と梃子比を考慮した誤差

以上より、誤差分散と梃子比を組み合わせて誤差の評価をするのが尤もらしい。それは、

$$t = \frac{z_{ek}}{\sqrt{1-h_{kk}}} \tag{139}$$

で与えられる。我々は、t に対して臨界値を与え、それを超えていれば外れ値として扱うことができる。これが位置によらずにそれぞれの点の誤差を評価している指標ととらえることができる。

4.3. 任意の説明変数種類数による回帰分析

ここでは、説明変数の種類数が m である任意の場合の解析を行う。これは、前節で求めた変数が 2 個の場合とほぼ同じ定式になる。行列形式で表現すると変数が 2 個の場合と同じである。したがって、導出方法は 2 変数の場合とほぼ同じである。ここでは、省略せずにその導出プロセスを書きしるす。

4.3.1. 回帰直線

我々は n セットのデータを母集団から取り、母集団の特性を予測するとする。

目的変数 y_k と m 個の説明変数 $x_{k1},x_{k2},\cdots,x_{km}$ は以下のように関連づけられるとする。

$$y_k = a_0 + a_1 x_{k1} + a_2 x_{k2} + \cdots + a_m x_{km} + e_k \tag{140}$$

と表現される。ここで n セットの各サンプルを添え字 k で表す。ここで $a_0,a_1,a_2,\cdots,a_m$ は母集団に関連する定数である。

係数 $a_0,a_1,a_2,\cdots,a_m$ を標本データから予想する。それは、母集団のそれとは異なる。したがって、標本データから予想したこれらの係数を $\hat{a}_0,\hat{a}_1,\hat{a}_2,\cdots,\hat{a}_m$ と表記する。

これらの標本データは

$$\begin{aligned}
y_1 &= \hat{a}_0 + \hat{a}_1 x_{11} + \hat{a}_2 x_{12} + \cdots + \hat{a}_m x_{1m} + e_1 \\
y_2 &= \hat{a}_0 + \hat{a}_1 x_{21} + \hat{a}_2 x_{22} + \cdots + \hat{a}_m x_{2m} + e_2 \\
&\cdots \\
y_n &= \hat{a}_0 + \hat{a}_1 x_{n1} + \hat{a}_2 x_{n2} + \cdots + \hat{a}_p x_{nm} + e_n
\end{aligned} \tag{141}$$

と表記される。回帰値 Y_k は

$$Y_k = \hat{a}_0 + \hat{a}_1 x_{k1} + \hat{a}_2 x_{k2} + \cdots + \hat{a}_m x_{km} \tag{142}$$

となる。誤差項 e_i は

$$e_i = y_i - Y_i = y_i - \left(\hat{a}_0 + \hat{a}_1 x_{i1} + \hat{a}_2 x_{i2} + \cdots + \hat{a}_m x_{im}\right) \tag{143}$$

となる。

この誤差項の平均は 0 で、かつそれは x_k と独立とする。つまり

$$E[e] = 0 \tag{144}$$

$$Cov\left[x_p, e\right] = 0 \tag{145}$$

とする。これらの誤差項と関連する仮定は後に検討され、誤差分散を最小にするという条件の下では成り立つことが示される。

誤差項 e_i の分散 $S_e^{(2)}$ は、以下のように与えられる。

$$S_e^{(2)} = \frac{1}{n}\sum_{k=1}^{n}\left[y_i - \left(\hat{a}_0 + \hat{a}_1 x_{k1} + \hat{a}_2 x_{k2} + \cdots + \hat{a}_m x_{km}\right)\right]^2 \tag{146}$$

我々は $S_e^{(2)}$ の値を最小にするように係数 $\hat{a}_0, \hat{a}_1, \hat{a}_2, \cdots, \hat{a}_m$ の値を決定する。

$S_e^{(2)}$ を $\hat{a}_0$ に関して偏微分し 0 と置くと

$$\frac{\partial S_e^{(2)}}{\partial \hat{a}_0} = -2\frac{1}{n}\sum_{k=1}^{n}\left[y_i - \left(\hat{a}_0 + \hat{a}_1 x_{k1} + \hat{a}_2 x_{k2} + \cdots + \hat{a}_m x_{km}\right)\right] = 0 \tag{147}$$

を得る。これから

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n}\left[y_k - \left(\hat{a}_0 + \hat{a}_1 x_{k1} + \hat{a}_2 x_{k2} + \cdots + \hat{a}_m x_{km}\right)\right] \\
&= \frac{1}{n}\sum_{k=1}^{n} y_k - \hat{a}_0 - \hat{a}_1\frac{1}{n}\sum_{k=1}^{n} x_{k1} - \hat{a}_2\frac{1}{n}\sum_{k=1}^{n} x_{k2} - \cdots - \hat{a}_m\frac{1}{n}\sum_{k=1}^{n} x_{km} \\
&= \bar{y} - \hat{a}_0 - \hat{a}_1\bar{x}_1 - \hat{a}_2\bar{x}_2 - \cdots - \hat{a}_m\bar{x}_m \\
&= 0
\end{aligned} \tag{148}$$

を得る。ただし、

$$\bar{x}_1 = \frac{1}{n}\sum_{k=1}^{n} x_{k1} \tag{149}$$

$$\bar{x}_2 = \frac{1}{n}\sum_{k=1}^{n} x_{k2} \tag{150}$$

・・・

$$\bar{x}_m = \frac{1}{n}\sum_{k=1}^{n} x_{km} \tag{151}$$

$$\bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k \tag{152}$$

である。

$S_e^{(2)}$ を $\hat{a}_1, \hat{a}_2, \cdots, \hat{a}_m$ で偏微分し 0 と置くと

$$\frac{\partial S_e^{(2)}}{\partial \hat{a}_1} = -2\frac{1}{n}\sum_{k=1}^{n} x_{k1}\left[y_k - \left(\hat{a}_0 + \hat{a}_1 x_{k1} + \hat{a}_2 x_{k2} + \cdots + \hat{a}_m x_{km}\right)\right] = 0 \qquad \text{eqa2(153)}$$

$$\frac{\partial S_e^{(2)}}{\partial \hat{a}_2} = -2\frac{1}{n}\sum_{k=1}^{n} x_{k2}\left[y_k - \left(\hat{a}_0 + \hat{a}_1 x_{k1} + \hat{a}_2 x_{k2} + \cdots + \hat{a}_m x_{km}\right)\right] = 0 \qquad \text{eqa3(154)}$$

・・・

$$\frac{\partial S_e^{(2)}}{\partial \hat{a}_m} = -2\frac{1}{n}\sum_{k=1}^{n} x_{km}\left[y_i - \left(\hat{a}_0 + \hat{a}_1 x_{k1} + \hat{a}_2 x_{k2} + \cdots + \hat{a}_m x_{km}\right)\right] = 0 \qquad \text{eqa3(155)}$$

となる Eq. (171)を Eq. (176)~(178)に代入して

$$\frac{1}{n}\sum_{k=1}^{n} x_{k1}\left[\left(y_k - \bar{y}\right) - \hat{a}_1\left(x_{k1} - \bar{x}_1\right) - \hat{a}_2\left(x_{k2} - \bar{x}_2\right) - \cdots - \hat{a}_m\left(x_{km} - \bar{x}_m\right)\right] = 0 \qquad (156)$$

$$\frac{1}{n}\sum_{k=1}^{n} x_{k2}\left[\left(y_k - \bar{y}\right) - \hat{a}_1\left(x_{k1} - \bar{x}_1\right) - \hat{a}_2\left(x_{k2} - \bar{x}_2\right) - \cdots - \hat{a}_m\left(x_{km} - \bar{x}_m\right)\right] = 0 \qquad (157)$$

・・・

$$\frac{1}{n}\sum_{k=1}^{n} x_{km}\left[\left(y_k - \bar{y}\right) - \hat{a}_1\left(x_{k1} - \bar{x}_1\right) - \hat{a}_2\left(x_{k2} - \bar{x}_2\right) - \cdots - \hat{a}_m\left(x_{km} - \bar{x}_m\right)\right] = 0 \qquad (158)$$

を得る。ここで、以下が成り立つことに注目する。

$$\frac{1}{n}\sum_{k=1}^{n} \bar{x}_1\left[\left(y_k - \bar{y}\right) - \hat{a}_1\left(x_{k1} - \bar{x}_1\right) - \hat{a}_2\left(x_{k2} - \bar{x}_2\right) - \cdots - \hat{a}_m\left(x_{km} - \bar{x}_m\right)\right] = 0 \qquad (159)$$

$$\frac{1}{n}\sum_{k=1}^{n} \bar{x}_2\left[\left(y_k - \bar{y}\right) - \hat{a}_1\left(x_{k1} - \bar{x}_1\right) - \hat{a}_2\left(x_{k2} - \bar{x}_2\right) - \cdots - \hat{a}_m\left(x_{km} - \bar{x}_m\right)\right] = 0 \qquad (160)$$

・・・

$$\frac{1}{n}\sum_{k=1}^{n} \bar{x}_m\left[\left(y_k - \bar{y}\right) - \hat{a}_1\left(x_{k1} - \bar{x}_1\right) - \hat{a}_2\left(x_{k2} - \bar{x}_2\right) - \cdots - \hat{a}_m\left(x_{km} - \bar{x}_m\right)\right] = 0 \qquad (161)$$

これらを利用して

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n}\left(x_{k1} - \bar{x}_1\right)\left[\left(y_k - \bar{y}\right) - \hat{a}_1\left(x_{k1} - \bar{x}_1\right) - \hat{a}_2\left(x_{k2} - \bar{x}_2\right) - \cdots - \hat{a}_m\left(x_{km} - \bar{x}_m\right)\right] \\
&= S_{1y}^{(2)} - \hat{a}_1 S_{11}^{(2)} - \hat{a}_2 S_{12}^{(2)} - \cdots - \hat{a}_m S_{1m}^{(2)} \\
&= 0
\end{aligned} \qquad (162)$$

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n}\left(x_{k2} - \bar{x}_2\right)\left[\left(y_k - \bar{y}\right) - \hat{a}_1\left(x_{k1} - \bar{x}_1\right) - \hat{a}_2\left(x_{k2} - \bar{x}_2\right) - \cdots - \hat{a}_m\left(x_{km} - \bar{x}_m\right)\right] \\
&= S_{2y}^{(2)} - \hat{a}_1 S_{21}^{(2)} - \hat{a}_2 S_{22}^{(2)} - \cdots - \hat{a}_m S_{2m}^{(2)} \\
&= 0
\end{aligned} \qquad (163)$$

・・・

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n}\left(x_{km} - \bar{x}_m\right)\left[\left(y_k - \bar{y}\right) - \hat{a}_1\left(x_{k1} - \bar{x}_1\right) - \hat{a}_2\left(x_{k2} - \bar{x}_2\right) - \cdots - \hat{a}_m\left(x_{km} - \bar{x}_m\right)\right] \\
&= S_{my}^{(2)} - \hat{a}_1 S_{m1}^{(2)} - \hat{a}_2 S_{m2}^{(2)} - \cdots - \hat{a}_m S_{mm}^{(2)} \\
&= 0
\end{aligned} \qquad (164)$$

ただし、以下の分散を定義している。

$$S_{ij}^{(2)} \equiv \frac{1}{n}\sum_{k=1}^{n}\left(x_{ki} - \bar{x}_i\right)\left(x_{kj} - \bar{x}_j\right) \qquad (165)$$

$$S_{yy}^{(2)} \equiv \frac{1}{n}\sum_{k=1}^{n}\left(y_k - \bar{y}\right)^2 \tag{166}$$

$$S_{iy}^{(2)} \equiv \frac{1}{n}\sum_{k=1}^{n}\left(x_{ki} - \bar{x}_i\right)\left(y_i - \bar{y}\right) \tag{167}$$

以上をまとめると

$$\begin{pmatrix} S_{11}^{(2)} & S_{12}^{(2)} & \cdots & S_{1m}^{(2)} \\ S_{21}^{(2)} & S_{22}^{(2)} & \cdots & S_{2m}^{(2)} \\ \cdots & \cdots & \ddots & \cdots \\ S_{m1}^{(2)} & S_{m2}^{(2)} & \cdots & S_{mm}^{(2)} \end{pmatrix} \begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \cdots \\ \hat{a}_m \end{pmatrix} = \begin{pmatrix} S_{1y}^{(2)} \\ S_{2y}^{(2)} \\ \cdots \\ S_{my}^{(2)} \end{pmatrix} \tag{168}$$

となる。したがって、各係数は

$$\begin{aligned} \begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \cdots \\ \hat{a}_m \end{pmatrix} &= \begin{pmatrix} S_{11}^{(2)} & S_{12}^{(2)} & \cdots & S_{1m}^{(2)} \\ S_{21}^{(2)} & S_{22}^{(2)} & \cdots & S_{2m}^{(2)} \\ \cdots & \cdots & \ddots & \cdots \\ S_{m1}^{(2)} & S_{m2}^{(2)} & \cdots & S_{mm}^{(2)} \end{pmatrix}^{-1} \begin{pmatrix} S_{1y}^{(2)} \\ S_{2y}^{(2)} \\ \cdots \\ S_{my}^{(2)} \end{pmatrix} \\ &= \begin{pmatrix} S^{11(2)} & S^{12(2)} & \cdots & S^{1m(2)} \\ S^{21(2)} & S^{22(2)} & \cdots & S^{2m(2)} \\ \cdots & \cdots & \ddots & \cdots \\ S^{m1(2)} & S^{m2(2)} & \cdots & S^{mm(2)} \end{pmatrix} \begin{pmatrix} S_{1y}^{(2)} \\ S_{2y}^{(2)} \\ \cdots \\ S_{my}^{(2)} \end{pmatrix} \end{aligned} \tag{169}$$

と求まる。ただし、

$$\begin{pmatrix} S^{11(2)} & S^{12(2)} & \cdots & S^{1m(2)} \\ S^{21(2)} & S^{22(2)} & \cdots & S^{2m(2)} \\ \cdots & \cdots & \ddots & \cdots \\ S^{m1(2)} & S^{m2(2)} & \cdots & S^{mm(2)} \end{pmatrix} = \begin{pmatrix} S_{11}^{(2)} & S_{12}^{(2)} & \cdots & S_{1m}^{(2)} \\ S_{21}^{(2)} & S_{22}^{(2)} & \cdots & S_{2m}^{(2)} \\ \cdots & \cdots & \ddots & \cdots \\ S_{m1}^{(2)} & S_{m2}^{(2)} & \cdots & S_{mm}^{(2)} \end{pmatrix}^{-1} \tag{170}$$

である。

$\hat{a}_1$，$\cdots$，$\hat{a}_m$ が求まれば、 $\hat{a}_0$ は以下のように求めることができる。

$$\hat{a}_0 = \bar{y} - \hat{a}_1\bar{x}_1 - \hat{a}_2\bar{x}_2 - \cdots - \hat{a}_m\bar{x}_m \tag{171}$$

4.3.2. 残差誤差e_kの評価

ここでは、仮定していた残差誤差e_iの特性に関して検討する。

e_kの平均は$\bar{e}$ と表記する。これは

$$
\begin{aligned}
\bar{e} &= \frac{1}{n}\sum_{k=1}^{n}\left(y_k - Y_k\right) \\
&= \frac{1}{n}\sum_{k=1}^{n}\left[y_k - \left(\hat{a}_0 + \hat{a}_1 x_{k1} + \hat{a}_2 x_{k2} + \cdots + \hat{a}_m x_{km}\right)\right] \\
&= \frac{1}{n}\sum_{k=1}^{n}\left[\left(y_k - \bar{y}\right) - \hat{a}_1\left(x_{k1} - \bar{x}_1\right) - \hat{a}_2\left(x_{k2} - \bar{x}_2\right) - \cdots - \hat{a}_m\left(x_{km} - \bar{x}_m\right)\right] \\
&= 0
\end{aligned}
\tag{172}
$$

なり、平均は仮定していた通りに 0 になる。

これと x_1 の共分散は

$$
\begin{aligned}
Cov\left[x_i, e\right] &= \frac{1}{n}\sum_{k=1}^{n}\left(x_{ki} - \bar{x}_k\right)e_k \\
&= \frac{1}{n}\sum_{k=1}^{n}\left(x_{ik} - \bar{x}_1\right)\left[y_k - \left(\hat{a}_0 + \hat{a}_1 x_k + \hat{a}_2 x_{k2} + \cdots + \hat{a}_m x_{km}\right)\right] \\
&= \frac{1}{n}\sum_{k=1}^{n}\left(x_{ki} - \bar{x}_i\right)\left[\left(y_k - \bar{y}\right) - \hat{a}_1\left(x_{k1} - \bar{x}_1\right) - \hat{a}_2\left(x_{k2} - \bar{x}_2\right) - \cdots - \hat{a}_m\left(x_{km} - \bar{x}_m\right)\right] \\
&= S_{iy}^{(2)} - \hat{a}_1 S_{i1}^{(2)} - \hat{a}_2 S_{i2}^{(2)} - \cdots - \hat{a}_m S_{im}^{(2)} \\
&= S_{iy}^{(2)} - S_{iy}^{(2)} \\
&= 0
\end{aligned}
\tag{173}
$$

したがって、共分散も 0 になる。

e_k の分散は

$$
\begin{aligned}
&S_e^{(2)}\\
&=\frac{e_1^{\,2}+e_2^{\,2}+\cdots+e_n^{\,2}}{n}\\
&=\frac{1}{n}\sum_{k=1}^{n}\left(y_k-\hat{a}_0-\hat{a}_1x_{k1}-\hat{a}_2x_{k2}-\cdots-\hat{a}_mx_{km}\right)^2\\
&=\frac{1}{n}\sum_{k=1}^{n}\left[\left(y_k-\bar{y}\right)-\hat{a}_1\left(x_{k1}-\bar{x}_1\right)-\hat{a}_2\left(x_{k2}-\bar{x}_2\right)-\cdots-\hat{a}_m\left(x_{km}-\bar{x}_m\right)\right]^2\\
&=\frac{1}{n}\sum_{k=1}^{n}\left(y_k-\bar{y}\right)^2+\hat{a}_1^{\,2}\frac{1}{n}\sum_{k=1}^{n}\left(x_{k1}-\bar{x}_1\right)^2+\hat{a}_2^{\,2}\frac{1}{n}\sum_{k=1}^{n}\left(x_{k2}-\bar{x}_2\right)^2+\cdots+\hat{a}_m^{\,2}\frac{1}{n}\sum_{k=1}^{n}\left(x_{km}-\bar{x}_m\right)^2\\
&-2\hat{a}_1\frac{1}{n}\sum_{k=1}^{n}\left(y_k-\bar{y}\right)\left(x_{k1}-\bar{x}_1\right)-2\hat{a}_2\frac{1}{n}\sum_{k=1}^{n}\left(y_k-\bar{y}\right)\left(x_{k2}-\bar{x}_2\right)-\cdots-2\hat{a}_m\frac{1}{n}\sum_{k=1}^{n}\left(y_k-\bar{y}\right)\left(x_{km}-\bar{x}_m\right)\\
&+2\hat{a}_1\hat{a}_2\frac{1}{n}\sum_{k=1}^{n}\left(x_{k1}-\bar{x}_1\right)\left(x_{k2}-\bar{x}_2\right)+2\hat{a}_1\hat{a}_3\frac{1}{n}\sum_{k=1}^{n}\left(x_{k1}-\bar{x}_1\right)\left(x_{k3}-\bar{x}_3\right)+\cdots+2\hat{a}_1\hat{a}_m\frac{1}{n}\sum_{k=1}^{n}\left(x_{k1}-\bar{x}_1\right)\left(x_{km}-\bar{x}_m\right)\\
&+2\hat{a}_2\hat{a}_3\frac{1}{n}\sum_{k=1}^{n}\left(x_{k2}-\bar{x}_2\right)\left(x_{k3}-\bar{x}_3\right)+2\hat{a}_2\hat{a}_4\frac{1}{n}\sum_{k=1}^{n}\left(x_{k2}-\bar{x}_2\right)\left(x_{k4}-\bar{x}_4\right)+\cdots+2\hat{a}_2\hat{a}_m\frac{1}{n}\sum_{k=1}^{n}\left(x_{k2}-\bar{x}_2\right)\left(x_{km}-\bar{x}_m\right)\\
&+\cdots\\
&+2\hat{a}_{m-1}\hat{a}_m\frac{1}{n}\sum_{k=1}^{n}\left(x_{k(m-1)}-\bar{x}_{(m-1)}\right)\left(x_{km}-\bar{x}_m\right)\\
&=S_{yy}^{\,(2)}+\hat{a}_1^{\,2}S_{11}^{\,(2)}+\hat{a}_2^{\,2}S_{22}^{\,(2)}+\cdots++\hat{a}_m^{\,2}S_{mm}^{\,(2)}\\
&-2\hat{a}_1S_{1y}^{\,(2)}-2\hat{a}_2S_{2y}^{\,(2)}-\cdots-2\hat{a}_mS_{my}^{\,(2)}\\
&+2\hat{a}_1\hat{a}_2S_{12}^{\,(2)}+2\hat{a}_1\hat{a}_3S_{13}^{\,(2)}+\cdots+2\hat{a}_1\hat{a}_mS_{1m}^{\,(2)}\\
&+2\hat{a}_2\hat{a}_3S_{23}^{\,(2)}+2\hat{a}_2\hat{a}_4S_{24}^{\,(2)}+\cdots+2\hat{a}_2\hat{a}_mS_{2m}^{\,(2)}\\
&+\cdots\\
&+2\hat{a}_{m-1}\hat{a}_mS_{(m-1)m}^{\,(2)}\\
&=S_{yy}^{\,(2)}\\
&+\hat{a}_1\left(\hat{a}_1S_{11}^{\,(2)}+\hat{a}_2S_{12}^{\,(2)}+\cdots+\hat{a}_mS_{1m}^{\,(2)}\right)+\hat{a}_2\left(\hat{a}_1S_{21}^{\,(2)}+\hat{a}_2S_{22}^{\,(2)}+\cdots+\hat{a}_mS_{2m}^{\,(2)}\right)+\cdots+\hat{a}_m\left(\hat{a}_1S_{m1}^{\,(2)}+\hat{a}_2Sm^{(2)}+\cdots+\hat{a}_mS_{mm}^{\,(2)}\right)\\
&-2\hat{a}_1S_{1y}^{\,(2)}-2\hat{a}_2S_{2y}^{\,(2)}-\cdots-2\hat{a}_mS_{my}^{\,(2)}\\
&=S_{yy}^{\,(2)}\\
&+\hat{a}_1S_{1y}^{\,(2)}+\hat{a}_2S_{2y}^{\,(2)}+\cdots+\hat{a}_mS_{my}^{\,(2)}\\
&-2\hat{a}_1S_{1y}^{\,(2)}-2\hat{a}_2S_{2y}^{\,(2)}-\cdots-2\hat{a}_mS_{my}^{\,(2)}\\
&=S_{yy}^{\,(2)}-\hat{a}_1S_{1y}^{\,(2)}-\hat{a}_2S_{2y}^{\,(2)}-\cdots-\hat{a}_mS_{my}^{\,(2)}
\end{aligned}
$$

(174)

となる。これは、0 になる必要はない。つまり、0 になることを仮定していない。

つまり、誤差分散 $S_e^{(2)}$ を最小になるように回帰直線の係数を定めると、誤差項に対する仮定は自動的に成り立っている。

4.3.3. 回帰分散による回帰直線の精度評価

回帰の精度は 2 変数の場合と同様になされる。

回帰直線の値と実際の目的変数の値の差の分散は$S_e^{(2)}$であり、これを最小にするように係数も定められた。したがって、回帰直線の精度は$S_e^{(2)}$と関連づける。一方、目的変数の分散$S_{yy}^{(2)}$はこの誤差の分散$S_e^{(2)}$と回帰直線の分散$S_R^{(2)}$の和からなることがこの場合です後に示される。つまり、

$$S_{yy}^{(2)} = S_e^{(2)} + S_R^{(2)} \tag{175}$$

である。ここで、回帰分散$S_R^{(2)}$は以下のように与えられる。

$$S_R^{(2)} = \frac{1}{n}\sum_{k=1}^{n}\left(Y_k - \bar{y}\right)^2 \tag{176}$$

である。

したがって、2 変数の場合と同様に、この$S_R^{(2)}$の$S_{yy}^{(2)}$に対する比、つまり決定係数を回帰直線の精度の指標にできる。決定係数は

$$R^2 = \frac{S_R^{(2)}}{S_{yy}^{(2)}} \tag{177}$$

で与えられる。

もしすべてのデータが回帰直線上にあれば、回帰の精度は完全であり、この場合は$R^2 = 1$となる。一方、回帰直線のデータと実際の目的変数の値の差がおおきければ、$S_R^{(2)}$に比べて$S_e^{(2)}$が大きくなり、つまり$S_{yy}^{(2)}$が大きくなりR^2は 1 より大分小さくなる。

上で利用した分散の関係を導く。

yの分散は以下のように導かれる。

$$\begin{aligned}
S_{yy}^{(2)} &= \frac{1}{n}\sum_{k=1}^{n}\left(y_k - \bar{y}\right)^2 \\
&= \frac{1}{n}\sum_{k=1}^{n}\left(y_k - Y_k + Y_k - \bar{y}\right)^2 \\
&= \frac{1}{n}\sum_{k=1}^{n}\left(y_k - Y_k\right)^2 + \frac{1}{n}\sum_{k=1}^{n}\left(Y_k - \bar{y}\right)^2 + 2\frac{1}{n}\sum_{k=1}^{n}\left(y_k - Y_k\right)\left(Y_k - \bar{y}\right) \\
&= \frac{1}{n}\sum_{k=1}^{n}e_k^2 + \frac{1}{n}\sum_{k=1}^{n}\left(Y_k - \bar{y}\right)^2 + 2\frac{1}{n}\sum_{k=1}^{n}e_k\left(\hat{a}_0 + \hat{a}_1 x_{k1} + \hat{a}_2 x_{k2} + \cdots + \hat{a}_m x_{km}\right) \\
&= \frac{1}{n}\sum_{k=1}^{n}e_k^2 + \frac{1}{n}\sum_{k=1}^{n}\left(Y_k - \bar{y}\right)^2 + 2\hat{a}_0\frac{1}{n}\sum_{k=1}^{n}e_k + 2\hat{a}_1\frac{1}{n}\sum_{k=1}^{n}e_k x_{k1} + 2\hat{a}_2\frac{1}{n}\sum_{k=1}^{n}e_k x_{k2} + \cdots + 2\hat{a}_m\frac{1}{n}\sum_{k=1}^{n}e_k x_{km} \\
&= S_e^{(2)} + S_R^{(2)}
\end{aligned}$$

(178)

ここで

$$\frac{1}{n}\sum_{k=1}^{n} e_k = \frac{1}{n}\sum_{k=1}^{n} e_k x_{k1} = \frac{1}{n}\sum_{k=1}^{n} e_k x_{k2} = \cdots = \frac{1}{n}\sum_{k=1}^{n} e_k x_{km} = 0 \tag{179}$$

を利用している。

先に、

$$\begin{aligned} S_e^{(2)} &= S_{yy}^{(2)} - \hat{a}_1 S_{1y}^{(2)} - \hat{a}_2 S_{2y}^{(2)} - \cdots - \hat{a}_m S_{my}^{(2)} \\ &= S_{yy}^{(2)} - \sum_{i=1}^{m} \hat{a}_i S_{iy}^{(2)} \end{aligned} \tag{180}$$

であったから、二つの関係式を比較すると

$$\begin{aligned} S_R^{(2)} &= \frac{1}{n}\sum_{k=1}^{n} \left(Y_k - \bar{y}\right)^2 \\ &= \sum_{i=1}^{m} \hat{a}_i S_{iy}^{(2)} \end{aligned} \tag{181}$$

となる。

4.3.4. 重相関係数による回帰直線の精度の評価

ここでは、回帰直線の精度を重相関係数から議論する。もし、回帰がうまくいっていれば、回帰データと実際の目的変数のデータの相関係数は 1 に近くなるはずである。したがって、この二つの変数 y_k と Y_k の相関係数を評価すればいい。この回帰相関係数を r_{mult} を置くと、それは

$$r_{mult} = \frac{\sum_{k=1}^{n} \left(y_k - \bar{y}\right)\left(Y_k - \bar{y}\right)}{\sqrt{\sum_{k=1}^{n} \left(y_k - \bar{y}\right)^2} \sqrt{\sum_{k=1}^{n} \left(Y_k - \bar{y}\right)^2}} \tag{182}$$

となる。

この分子を評価する。まず、誤差に関する分散を変形すると

$$\begin{aligned} S_e^{(2)} &= \frac{1}{n}\sum_{k=1}^{n} \left(y_k - Y_k\right)^2 \\ &= \frac{1}{n}\sum_{k=1}^{n} \left[\left(y_k - \bar{y}\right) - \left(Y_k - \bar{y}\right)\right]^2 \\ &= \frac{1}{n}\sum_{k=1}^{n} \left(y_k - \bar{y}\right)^2 + \frac{1}{n}\sum_{k=1}^{n} \left(Y_k - \bar{y}\right)^2 - 2\frac{1}{n}\sum_{k=1}^{n} \left(y_k - \bar{y}\right)\left(Y_k - \bar{y}\right) \\ &= S_{yy}^{(2)} + S_R^{(2)} - 2\frac{1}{n}\sum_{k=1}^{n} \left(y_k - \bar{y}\right)\left(Y_k - \bar{y}\right) \end{aligned} \tag{183}$$

となり、重相関係数の分子の項が出てくる。これは

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n}\left(y_k-\bar{y}\right)\left(Y_k-\bar{y}\right)&=\frac{1}{2}\left(S_{yy}^{(2)}+S_R^{(2)}-S_e^{(2)}\right)\\&=\frac{1}{2}\left[S_{yy}^{(2)}+S_R^{(2)}-\left(S_{yy}^{(2)}-S_R^{(2)}\right)\right]\\&=S_R^{(2)}\end{aligned}\tag{184}$$

となる。したがって、重相関係数は

$$\begin{aligned}r_{mult}&=\frac{S_R^{(2)}}{\sqrt{S_{yy}^{(2)}}\sqrt{S_R^{(2)}}}\\&=\sqrt{\frac{S_R^{(2)}}{S_{yy}^{(2)}}}\end{aligned}\tag{185}$$

となる。したがって、重相関係数 r_{mult} の 2 乗が決定係数となる。つまり、両者はほぼ同じことをしている。式で表すと

$$r_{mult}=\sqrt{R^2}\tag{186}$$

となる。

4.3.5. 自由度調整済決定係数

回帰の精度はこれまで示してきたように決定係数または重相関係数で評価できる。この両者はほぼ同じことを評価している。この決定係数のさらに精度を上げる。決定係数は回帰分散と全体のデータの分散の比であるが、変形すると

$$R^2=\frac{S_R^{(2)}}{S_{yy}^{(2)}}=\frac{S_{yy}^{(2)}-S_e^{(2)}}{S_{yy}^{(2)}}=1-\frac{S_e^{(2)}}{S_{yy}^{(2)}}\tag{187}$$

となる。つまり、同じことであるが、誤差の分散と目的変数の分散の比が小さいと精度は高い、と評価される。しかし、この分散では自由度を考慮していない。ここでは、それらの分散の自由度を評価し、それを評価に取り入れる。

しかし、この分散では自由度を考慮していない。ここでは、それらの分散の自由度を評価し、それを評価に取り入れる。

まず、$S_{yy}^{(2)}$ の自由度 ϕ_r を考える。これは n 個のデータからなり、分散を評価するうえでその平均を利用している。つまり、その自由度は

$$\phi_r=n-1\tag{188}$$

である。

次に回帰分散 $S_R^{(2)}$ の自由度 ϕ_R を考える。これは m 個の変数 $\hat{a}_1,\hat{a}_2,\cdots,\hat{a}_m$ で評価される。つまり、その自由度は

$$\phi_R = m \tag{189}$$

である。

最後に誤差分散 $S_e^{(2)}$ の自由度 ϕ_e を考える。これは n 個の目的変数データと、 $m+1$ 個の係数で定まる回帰式の差である。つまり、その自由度は

$$\phi_e = n-(m+1) \tag{190}$$

である。

各分散の関係は

$$S_{yy}^{(2)} = S_R^{(2)} + S_e^{(2)} \tag{191}$$

であり、これは自由度の関係の方程式と直接関連する。すなわち

$$\phi_T = \phi_R + \phi_e \tag{192}$$

である。この関係は常に成り立つ。すなわち、変数の方程式と変数の自由度の方程式は常に 1 対 1 に対応する。

したがって、決定係数はその自由度を考慮した

$$R^{*2} = 1-\frac{\dfrac{n}{\phi_e}S_e^{(2)}}{\dfrac{n}{\phi_T}S_{yy}^{(2)}} \tag{193}$$

で評価される。こちらのほうが、より定量的に意味のある回帰の評価式である。これを調整済決定係数と呼ぶ。実際の値を代入すると

$$\begin{aligned} R^{*2} &= 1-\frac{\dfrac{n}{\phi_e}S_e^{(2)}}{\dfrac{n}{\phi_T}S_{yy}^{(2)}} \\ &= 1-\frac{\dfrac{n}{n-(m+1)}S_e^{(2)}}{\dfrac{n}{n-1}S_{yy}^{(2)}} \\ &= 1-\frac{n-1}{n-(m+1)}\frac{S_e^{(2)}}{S_{yy}^{(2)}} \\ &= 1-\frac{1}{1-\dfrac{m}{n-1}}\frac{S_e^{(2)}}{S_{yy}^{(2)}} \end{aligned} \tag{194}$$

となる。つまり、決定係数よりも小さくなる。

4.3.6. F 値による回帰精度の評価

回帰の精度は回帰分散と残渣誤差分散の比から決定できる。この比、

$$F = \frac{s_R^{(2)}}{s_e^{(2)}} \tag{195}$$

が大きければ、回帰の精度はいいといえる。ただし、

$$s_R^{(2)} = \frac{nS_R^{(2)}}{m} \tag{196}$$

$$s_e^{(2)} = \frac{nS_e^{(2)}}{n-(m+1)} \tag{197}$$

である。これと F 値の推定確率 P の場合の臨界値 $F_p\left(m,n-(m+1)\right)$ と比較し、この F 値が F_P よりも大きければ回帰の精度は高い、と判断できる。つまり以下の式で表現される。

$$\begin{cases} F \le F_p\left(m,n-(m+1)\right) \Rightarrow invalid \\ F > F_p\left(m,n-(m+1)\right) \Rightarrow valid \end{cases} \tag{198}$$

Microsoft の Excel はよく利用される。その出力に分散分析表がある。その中で有意 F 値がある。この有意 F 値は上の定義の F 値ではなく、以下のように定義される。

まず、観測された分散比が表示される。それは、この本の

$$F = \frac{s_R^{(2)}}{s_e^{(2)}} \tag{199}$$

のことである。その次の有意 F 値とは、それに対する P 点、すなわち

$$\text{有意}F\text{値} = \int_{x=\frac{s_r^{(2)}}{s_e^{(2)}}}^{\infty} \frac{n_1^{\frac{n_1}{2}} n_2^{\frac{n_2}{2}} x^{\frac{n_1-2}{2}}}{B\left(\frac{n_1}{2},\frac{n_2}{2}\right)\left(n_1 x+n_2\right)^{\frac{n_1+n_2}{2}}} dx \tag{200}$$

で定義される。ただし、

$$\begin{cases} n_1 = m \\ n_2 = n-(m+1) \end{cases} \tag{201}$$

である。つまり、Microsoft の Excel における有意 F 値は F 関数ではなく、観測される分散比つまり F 値に対する P 点である。すなわち、これが設定した推定確率 0.95 に関する確率

$1-P=1-0.95=0.05$よりも小さければ回帰はうまくいっているとみなす。

このF値を用いるほうが、推定確率と連動させて回帰の精度が十分かそうでないかを判断できる。

4.3.7. 回帰直線の係数の変動

標本データを変えるたびに取得するデータ値は異なり、それから導き出す回帰直線も異なったものになる。つまり、回帰直線の係数$\hat{a}_0,\hat{a}_1,\hat{a}_2,\cdots,\hat{a}_m$は抽出されたデータに応じて変動する。ここでは、$\hat{a}_0,\hat{a}_1,\hat{a}_2,\cdots,\hat{a}_m$の平均と分散および共分散を検討する。

母集団の回帰係数を仮定して、目的変数は以下のように表現されると仮定する。

$$y_k=a_0+a_1x_{k1}+a_2x_{k2}+\cdots+a_mx_{km}+e_k \quad for\ k=1,2,\cdots,n \tag{202}$$

ここで$(x_{k1},x_{k2},\cdots,x_{km})$は与えられたデータであり、係数$a_0$，$a_1$，$a_2$，$\cdots$,$a_m$は母集団のものである。これらの係数は確定した値であり、変動することはない。e_iは他の変数とは独立で正規分布$N\left[0,\sigma^{(2)}\right]$に従うとする。この方程式の中で$e_k$だけが確率変数である。

ここでは、2変数の場合の解析を拡張する。

$k=1,2,\cdots,n$についてのデータ$(x_{k1},x_{k2},\cdots,x_{km},y_i)$を取得したとする。その係数は

$$\begin{aligned}\begin{pmatrix}\hat{a}_1\\ \hat{a}_2\\ \vdots\\ \hat{a}_m\end{pmatrix}&=\begin{pmatrix}S^{11(2)} & S^{12(2)} & \cdots & S^{1m(2)}\\ S^{21(2)} & S^{22(2)} & \cdots & S^{1m(2)}\\ \vdots & \vdots & \ddots & \vdots\\ S^{m1(2)} & S^{m2(2)} & \cdots & S^{mm(2)}\end{pmatrix}\begin{pmatrix}S_{1y}^{(2)}\\ S_{2y}^{(2)}\\ \vdots\\ S_{my}^{(2)}\end{pmatrix}\\ &=\begin{pmatrix}S^{11(2)}S_{1y}^{(2)}+S^{12(2)}S_{2y}^{(2)}+\cdots+S^{1m(2)}S_{my}^{(2)}\\ S^{21(2)}S_{1y}^{(2)}+S^{22(2)}S_{2y}^{(2)}+\cdots+S^{2m(2)}S_{my}^{(2)}\\ \cdots\\ S^{m1(2)}S_{1y}^{(2)}+S^{m2(2)}S_{2y}^{(2)}+\cdots+S^{mm(2)}S_{my}^{(2)}\end{pmatrix}\end{aligned} \tag{203}$$

となる。それぞれは、式を展開して

$$\begin{aligned}\hat{a}_1&=S^{11(2)}S_{1y}^{(2)}+S^{12(2)}S_{2y}^{(2)}+\cdots+S^{1m(2)}S_{my}^{(2)}\\ &=\sum_{p=1}^{m}S^{1p(2)}S_{py}^{(2)}\\ &=\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{m}S^{1p(2)}\left(x_{kp}-\bar{x}_p\right)\right]y_k\\ &=\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{m}S^{1p(2)}\left(x_{kp}-\bar{x}_p\right)\right]\left(a_0+a_1x_{k1}+a_2x_{k2}+\cdots+a_mx_{km}+e_k\right)\end{aligned} \tag{204}$$

$$
\begin{aligned}
\hat{a}_2 &= S^{21(2)}S_{1y}^{(2)} + S^{22(2)}S_{2y}^{(2)} + \cdots + S^{2m(2)}S_{my}^{(2)} \\
&= \sum_{p=1}^{m} S^{2p(2)}S_{py}^{(2)} \\
&= \frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{m} S^{2p(2)}\left(x_{kp} - \overline{x}_p\right)\right] y_k \\
&= \frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{m} S^{2p(2)}\left(x_{kp} - \overline{x}_p\right)\right]\left(a_0 + a_1 x_{k1} + a_2 x_{k2} + \cdots + a_m x_{km} + e_k\right)
\end{aligned}
\tag{205}
$$

$$
\begin{aligned}
\hat{a}_0 &= \overline{y} - \left(\hat{a}_1\overline{x}_1 + \hat{a}_2\overline{x}_2 + \cdots + \hat{a}_m\overline{x}_m\right) \\
&= \frac{1}{n}\sum_{k=1}^{n} y_k - \frac{1}{n}\sum_{k=1}^{n}\left[\overline{x}_1\sum_{p=1}^{m} S^{1p(2)}\left(x_{kp} - \overline{x}_p\right)\right] y_k - \frac{1}{n}\sum_{k=1}^{n}\left[\overline{x}_2\sum_{p=1}^{m} S^{2p(2)}\left(x_{kp} - \overline{x}_p\right)\right] y_k - \cdots - \frac{1}{n}\sum_{k=1}^{n}\left[\overline{x}_m\sum_{p=1}^{m} S^{mp(2)}\left(x_{kp} - \overline{x}_p\right)\right] y_k \\
&= \frac{1}{n}\sum_{k=1}^{n}\left\{1 - \overline{x}_1\sum_{p=1}^{m} S^{1p(2)}\left(x_{kp} - \overline{x}_p\right) - \overline{x}_2\sum_{p=1}^{m} S^{2p(2)}\left(x_{kp} - \overline{x}_p\right) - \cdots - \overline{x}_m\sum_{p=1}^{m} S^{mp(2)}\left(x_{kp} - \overline{x}_p\right)\right\} y_k \\
&= \frac{1}{n}\sum_{k=1}^{n}\left\{1 - \sum_{j=1}^{m}\overline{x}_j\sum_{p=1}^{m} S^{jp(2)}\left(x_{kp} - \overline{x}_p\right)\right\} y_k \\
&= \frac{1}{n}\sum_{k=1}^{n}\left\{1 - \sum_{j=1}^{m}\overline{x}_j\sum_{p=1}^{m} S^{jp(2)}\left(x_{kp} - \overline{x}_p\right)\right\}\left(a_0 + a_1 x_{k1} + a_2 x_2 + \cdots + a_m x_{km} + e_k\right)
\end{aligned}
$$

(206)

となる。ここで、再び e_k だけが確率変数であることに注目する。そして、期待値は以下のように評価される。

$$
\begin{aligned}
E\left[\hat{a}_1\right] &= E\left[\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{m}S^{1p(2)}\left(x_{kp}-\bar{x}_p\right)\right]\left(a_0+a_1x_{k1}+a_2x_{k2}+\cdots+a_mx_{km}+e_k\right)\right] \\
&= a_0\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{m}S^{1p(2)}\left(x_{kp}-\bar{x}_p\right)\right] \\
&\quad +a_1\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{m}S^{1p(2)}\left(x_{kp}-\bar{x}_p\right)\right]x_{k1} \\
&\quad +a_2\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{m}S^{1p(2)}\left(x_{kp}-\bar{x}_p\right)\right]x_{k2} \\
&\quad +\cdots \\
&\quad +a_m\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{m}S^{1p(2)}\left(x_{kp}-\bar{x}_p\right)\right]x_{km} \\
&\quad +E\left[\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{m}S^{1p(2)}\left(x_{kp}-\bar{x}_p\right)\right]e_k\right] \\
&= a_1\sum_{p=1}^{m}S^{1p(2)}S_{p1}^{(2)}+a_2\sum_{p=1}^{m}S^{1p(2)}S_{p2}^{(2)}+\cdots+a_m\sum_{p=1}^{m}S^{1p(2)}S_{pm}^{(2)} \\
&= a_1
\end{aligned}
\tag{207}
$$

ところで

$$
\sum_{p=1}^{m}S^{ip(2)}S_{pj}^{(2)}=\delta_{ij} \tag{208}
$$

を利用している。

$$
\begin{aligned}
E\left[\hat{a}_2\right] &= \frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{m}S^{2p(2)}\left(x_{kp}-\bar{x}_p\right)\right]\left(a_0+a_1x_{k1}+a_2x_{k2}+\cdots+a_mx_{km}+e_k\right) \\
&= a_0\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{m}S^{2p(2)}\left(x_{kp}-\bar{x}_p\right)\right] \\
&\quad +a_1\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{m}S^{2p(2)}\left(x_{kp}-\bar{x}_p\right)\right]x_{k1} \\
&\quad +a_2\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{m}S^{2p(2)}\left(x_{kp}-\bar{x}_p\right)\right]x_{k2} \\
&\quad +\cdots \\
&\quad +a_m\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{m}S^{2p(2)}\left(x_{kp}-\bar{x}_p\right)\right]x_{km} \\
&\quad +E\left[\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{p=1}^{m}S^{2p(2)}\left(x_{kp}-\bar{x}_p\right)\right]e_k\right] \\
&= a_1\sum_{p=1}^{2}S^{2p(2)}S_{p1}^{(2)}+a_2\sum_{p=1}^{2}S^{2p(2)}S_{p2}^{(2)}+\cdots+a_m\sum_{p=1}^{2}S^{2p(2)}S_{pm}^{(2)} \\
&= a_2
\end{aligned}
\tag{209}
$$

$$
\begin{aligned}
E\left[\hat{a}_0\right] &= E\left[\frac{1}{n}\sum_{k=1}^{n}\left\{1-\sum_{j=1}^{m}\bar{x}_j\sum_{p=1}^{m}S^{jp(2)}\left(x_{kp}-\bar{x}_p\right)\right\}\left(a_0+a_1x_{k1}+a_2x_{k2}+\cdots+a_mx_{km}+e_k\right)\right] \\
&= a_0\frac{1}{n}\sum_{k=1}^{n}\left[1-\sum_{j=1}^{m}\bar{x}_j\sum_{p=1}^{m}S^{jp(2)}\left(x_{kp}-\bar{x}_p\right)\right] \\
&+a_1\frac{1}{n}\sum_{k=1}^{n}\left[1-\sum_{j=1}^{m}\bar{x}_j\sum_{p=1}^{m}S^{jp(2)}\left(x_{kp}-\bar{x}_p\right)\right]x_{k1} \\
&+a_2\frac{1}{n}\sum_{k=1}^{n}\left[1-\sum_{j=1}^{m}\bar{x}_j\sum_{p=1}^{m}S^{jp(2)}\left(x_{kp}-\bar{x}_p\right)\right]x_{k2} \\
&+\cdots \\
&+a_m\frac{1}{n}\sum_{k=1}^{n}\left[1-\sum_{j=1}^{m}\bar{x}_j\sum_{p=1}^{m}S^{jp(2)}\left(x_{kp}-\bar{x}_p\right)\right]x_{km} \\
&+E\left[\frac{1}{n}\sum_{k=1}^{n}\left[1-\sum_{j=1}^{m}\bar{x}_j\sum_{p=1}^{m}S^{jp(2)}\left(x_{kp}-\bar{x}_p\right)\right]e_k\right]
\end{aligned}
\tag{210}
$$

ここで、最初の項を考える。

$$
\begin{aligned}
&a_0\frac{1}{n}\sum_{k=1}^{n}\left[1-\sum_{j=1}^{m}\bar{x}_j\sum_{p=1}^{m}S^{jp(2)}\left(x_{kp}-\bar{x}_p\right)\right] \\
&= a_0-a_0\sum_{k=1}^{n}\left[\bar{x}_1\sum_{p=1}^{m}S^{1p(2)}\left(x_{kp}-\bar{x}_p\right)\right]-a_0\sum_{k=1}^{n}\left[\bar{x}_2\sum_{p=1}^{m}S^{2p(2)}\left(x_{kp}-\bar{x}_p\right)\right]-\cdots-a_0\sum_{k=1}^{n}\left[\bar{x}_m\sum_{p=1}^{m}S^{2p(2)}\left(x_{kp}-\bar{x}_p\right)\right] \\
&= a_0-a_0\left[\bar{x}_1\sum_{p=1}^{2}S^{1p(2)}\sum_{k=1}^{n}\left(x_{kp}-\bar{x}_p\right)\right]-a_0\left[\bar{x}_2\sum_{p=1}^{2}S^{2p(2)}\sum_{k=1}^{n}\left(x_{kp}-\bar{x}_p\right)\right]-\cdots-a_0\left[\bar{x}_m\sum_{p=1}^{2}S^{2p(2)}\sum_{k=1}^{n}\left(x_{kp}-\bar{x}_p\right)\right] \\
&= a_0
\end{aligned}
\tag{211}
$$

二番目の項を考える。

$$
\begin{aligned}
&a_1\frac{1}{n}\sum_{k=1}^{n}\left[1-\sum_{j=1}^{m}\bar{x}_j\sum_{p=1}^{m}S^{jp(2)}\left(x_{kp}-\bar{x}_p\right)\right]x_{k1} \\
&= a_1\bar{x}_1-a_1\frac{1}{n}\sum_{k=1}^{n}\left[\bar{x}_1\sum_{p=1}^{m}S^{1p(2)}\left(x_{kp}-\bar{x}_p\right)\right]x_{k1}-a_1\frac{1}{n}\sum_{k=1}^{n}\left[\bar{x}_2\sum_{p=1}^{m}S^{2p(2)}\left(x_{kp}-\bar{x}_p\right)\right]x_{k1}-\cdots-a_1\frac{1}{n}\sum_{k=1}^{n}\left[\bar{x}_m\sum_{p=1}^{m}S^{2p(2)}\left(x_{kp}-\bar{x}_p\right)\right]x_{k1} \\
&= a_1\bar{x}_1-a_1\bar{x}_1\sum_{p=1}^{m}S^{1p(2)}S_{p1}^{(2)}-a_1\bar{x}_2\sum_{p=1}^{m}S^{2p(2)}S_{p1}^{(2)}-\cdots-a_1\bar{x}_m\sum_{p=1}^{m}S^{2p(2)}S_{p1}^{(2)} \\
&= a_1\bar{x}_1-a_1\bar{x}_1S^{11(2)}S_{11}^{(2)} \\
&= 0
\end{aligned}
\tag{212}
$$

第 3 項目を考える。

$$
\begin{aligned}
&a_2 \frac{1}{n}\sum_{k=1}^{n}\left[1-\sum_{j=1}^{m}\bar{x}_j\sum_{p=1}^{m}S^{jp(2)}\left(x_{kp}-\bar{x}_p\right)\right]x_{k2}\\
&=a_2\bar{x}_2-a_2\bar{x}_1\sum_{p=1}^{m}S^{1p(2)}\sum_{k=1}^{n}\left(x_{kp}-\bar{x}_p\right)x_{k2}-a_2\bar{x}_2\sum_{p=1}^{m}S^{2p(2)}\sum_{k=1}^{n}\left(x_{kp}-\bar{x}_p\right)x_{k2}-\cdots-a_2\bar{x}_m\sum_{p=1}^{m}S^{2p(2)}\sum_{k=1}^{n}\left(x_{kp}-\bar{x}_p\right)x_{k2}\\
&=a_2\bar{x}_2-a_2\bar{x}_1\sum_{p=1}^{m}S^{1p(2)}S_{p2}^{(2)}-a_2\bar{x}_2\sum_{p=1}^{m}S^{2p(2)}S_{p2}^{(2)}-\cdots--a_2\bar{x}_m\sum_{p=1}^{m}S^{2p(2)}S_{p2}^{(2)}\\
&=a_2\bar{x}_2-a_2\bar{x}_2S^{22(2)}S_{22}^{(2)}\\
&=a_2\bar{x}_2-a_2\bar{x}_2\\
&=0
\end{aligned}
\tag{213}
$$

他も同様に 0 になる。

最後の項を考える。

$$
\begin{aligned}
&E\left[\frac{1}{n}\sum_{k=1}^{n}\left[1-\sum_{j=1}^{m}\bar{x}_j\sum_{p=1}^{m}S^{jp(2)}\left(x_{kp}-\bar{x}_p\right)e_k\right]\right]\\
&=E\left[\frac{1}{n}\sum_{k=1}^{n}e_k\right]-\sum_{j=1}^{m}\bar{x}_j\sum_{p=1}^{m}S^{jp(2)}E\left[\frac{1}{n}\sum_{k=1}^{n}x_{kp}e_k\right]+\sum_{j=1}^{m}\bar{x}_j\sum_{p=1}^{m}S^{jp(2)}\bar{x}_pE\left[\frac{1}{n}\sum_{k=1}^{n}e_k\right]\\
&=0
\end{aligned}
\tag{214}
$$

したがって、以下を得る。

$$E\left[\hat{a}_0\right]=a_0 \tag{215}$$

次に分散を評価する。それは以下のようになる。

$$
\begin{aligned}
V\left[\hat{a}_1\right] &= \sum_{k=1}^{n}\left[\frac{1}{n}\sum_{p=1}^{m} S^{1p(2)}\left(x_{kp}-\bar{x}_p\right)\right]^2 V\left[e_k\right] \\
&= \frac{1}{n}\frac{1}{n}\sum_{k=1}^{n}\left[S^{11(2)}\left(x_{k1}-\bar{x}_1\right)+S^{12(2)}\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{1m(2)}\left(x_{km}-\bar{x}_m\right)\right]^2 V\left[e_k\right] \\
&= \frac{1}{n}\frac{1}{n}\sum_{k=1}^{n}\left[\begin{array}{l}
\left(S^{11(2)}\right)^2\left(x_{k1}-\bar{x}_1\right)^2+\left(S^{12(2)}\right)^2\left(x_{k2}-\bar{x}_2\right)^2+\cdots+\left(S^{1m(2)}\right)^2\left(x_{km}-\bar{x}_m\right)^2 \\
+2S^{11(2)}S^{12(2)}\left(x_{k1}-\bar{x}_1\right)\left(x_{k2}-\bar{x}_2\right)+2S^{11(2)}S^{12(2)}\left(x_{k1}-\bar{x}_1\right)\left(x_{k2}-\bar{x}_2\right)+\cdots+2S^{11(2)}S^{1m(2)}\left(x_{k1}-\bar{x}_1\right)\left(x_{km}-\bar{x}_m\right) \\
+2S^{12(2)}S^{13(2)}\left(x_{22}-\bar{x}_2\right)\left(x_{k3}-\bar{x}_3\right)+2S^{12(2)}S^{14(2)}\left(x_{k2}-\bar{x}_2\right)\left(x_{k4}-\bar{x}_4\right)+\cdots+2S^{12(2)}S^{1m(2)}\left(x_{k2}-\bar{x}_2\right)\left(x_{km}-\bar{x}_m\right) \\
+\cdots \\
+2S^{1(m-1)(2)}S^{1m(2)}\left(x_{k(m-1)}-\bar{x}_{(m-1)}\right)\left(x_{km}-\bar{x}_m\right)
\end{array}\right] \\
&= \frac{1}{n}\left[\left(S^{11(2)}\right)^2 S_{11}^{(2)}+2S^{12(2)}S^{11(2)}S_{12}^{(2)}+\left(S^{12(2)}\right)^2 S_{22}^{(2)}\right]s_e^{(2)} \\
&= \frac{1}{n}\left[\begin{array}{l}
S^{11(2)}\left[S^{11(2)}S_{11}^{(2)}+S^{12(2)}S_{21}^{(2)}+\cdots+S^{1m(2)}S_{m1}^{(2)}\right] \\
+S^{12(2)}\left[S^{11(2)}S_{12}^{(2)}+S^{12(2)}S_{22}^{(2)}+\cdots+S^{1m(2)}S_{m2}^{(2)}\right] \\
+\cdots \\
+S^{1m(2)}\left[S^{11(2)}S_{1m}^{(2)}+S^{12(2)}S_{2m}^{(2)}+\cdots+S^{1m(2)}S_{mm}^{(2)}\right]
\end{array}\right]s_e^{(2)} \\
&= \frac{S^{11(2)}}{n}s_e^{(2)}
\end{aligned}
$$

(216)

ここで、

$$
S^{i1(2)}S_{1j}^{(2)}+S^{i2(2)}S_{2j}^{(2)}+\cdots+S^{im(2)}S_{mj}^{(2)}=\delta_{ij} \tag{217}
$$

を利用している。

$$
\begin{aligned}
V\left[\hat{a}_2\right] &= \sum_{k=1}^{n}\left[\frac{1}{n}\sum_{p=1}^{m} S^{2p(2)}\left(x_{kp}-\bar{x}_p\right)\right]^2 V[e] \\
&= \frac{1}{n}\frac{1}{n}\sum_{k=1}^{n}\left[S^{21(2)}\left(x_{k1}-\bar{x}_1\right)+S^{22(2)}\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{2m(2)}\left(x_{km}-\bar{x}_m\right)\right]^2 s_e^{(2)} \\
&= \frac{1}{n}\frac{1}{n}\sum_{k=1}^{n}\left[\begin{array}{l}
\left(S^{21(2)}\right)^2\left(x_{k1}-\bar{x}_1\right)^2+\left(S^{22(2)}\right)^2\left(x_{k2}-\bar{x}_2\right)^2+\cdots+\left(S^{2m(2)}\right)^2\left(x_{km}-\bar{x}_m\right)^2 \\
+2S^{21(2)}S^{22(2)}\left(x_{k1}-\bar{x}_1\right)\left(x_{k2}-\bar{x}_2\right)+2S^{21(2)}S^{23(2)}\left(x_{k1}-\bar{x}_1\right)\left(x_{k3}-\bar{x}_3\right)+\cdots+2S^{21(2)}S^{2m(2)}\left(x_{k1}-\bar{x}_1\right)\left(x_{km}-\bar{x}_m\right) \\
+2S^{22(2)}S^{23(2)}\left(x_{k2}-\bar{x}_2\right)\left(x_{k3}-\bar{x}_3\right)+2S^{22(2)}S^{24(2)}\left(x_{k2}-\bar{x}_2\right)\left(x_{k4}-\bar{x}_4\right)+\cdots+2S^{22(2)}S^{2m(2)}\left(x_{k2}-\bar{x}_2\right)\left(x_{km}-\bar{x}_m\right) \\
+\cdots \\
+2S^{2(m-1)(2)}S^{2m(2)}\left(x_{k(m-1)}-\bar{x}_{(m-1)}\right)\left(x_{km}-\bar{x}_m\right)
\end{array}\right] \\
&= \frac{1}{n}\left[\begin{array}{l}
S^{21(2)}\left[S^{21(2)}S_{11}^{(2)}+S^{22(2)}S_{21}^{(2)}+\cdots+S^{2m(2)}S_{m1}^{(2)}\right] \\
+S^{22(2)}\left[S^{21(2)}S_{12}^{(2)}+S^{22(2)}S_{22}^{(2)}+\cdots+S^{2m(2)}S_{m2}^{(2)}\right] \\
+\cdots \\
+S^{2m(2)}\left[S^{21(2)}S_{1m}^{(2)}+S^{22(2)}S_{2m}^{(2)}+\cdots+S^{2m(2)}S_{mm}^{(2)}\right]
\end{array}\right] s_e^{(2)} \\
&= \frac{S^{22(2)}}{n} s_e^{(2)}
\end{aligned}
$$

(218)

他の係数の同様である。

すなわち

$$
V\left[\hat{a}_p\right] = \frac{S^{pp(2)}}{n} s_e^{(2)} \tag{219}
$$

である。ただし、 $p \neq 0$ である。

最後に $\hat{a}_0$ の分散を考える。

$$
\begin{aligned}
V\left[\hat{a}_0\right] &= V\left[\frac{1}{n}\sum_{k=1}^{n}\left\{1-\sum_{j=1}^{m}\bar{x}_j\sum_{p=1}^{m}S^{jp(2)}\left(x_{kp}-\bar{x}_p\right)\right\}\left(a_0+a_1x_{k1}+a_2x_{k2}+\cdots+a_mx_{km}+e_k\right)\right] \\
&= \left[\sum_{i=1}^{n}\left[\frac{1}{n}\left\{1-\sum_{j=1}^{m}\bar{x}_j\sum_{p=1}^{m}S^{jp(2)}\left(x_{ip}-\bar{x}_p\right)\right\}\right]^2\right]V\left[e\right] \\
&= \frac{1}{n}\frac{1}{n}\left[\sum_{k=1}^{n}\left[\left\{1-\sum_{j=1}^{m}\bar{x}_j\left[S^{j1(2)}\left(x_{k1}-\bar{x}_1\right)+S^{j2(2)}\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{jm(2)}\left(x_{km}-\bar{x}_m\right)\right]\right\}\right]^2\right]V\left[e\right] \\
&= \frac{1}{n}\frac{1}{n}\left[\sum_{k=1}^{n}\left[1-\left\{\begin{array}{l}\bar{x}_1\left[S^{11(2)}\left(x_{k1}-\bar{x}_1\right)+S^{12(2)}\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{1m(2)}\left(x_{km}-\bar{x}_m\right)\right] \\ +\bar{x}_2\left[S^{21(2)}\left(x_{k1}-\bar{x}_1\right)+S^{22(2)}\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{2m(2)}\left(x_{km}-\bar{x}_m\right)\right] \\ +\cdots \\ +\bar{x}_m\left[S^{m1(2)}\left(x_{k1}-\bar{x}_1\right)+S^{m2(2)}\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{mm(2)}\left(x_{km}-\bar{x}_m\right)\right]\end{array}\right\}\right]^2\right]s_e^{(2)} \\
&= \frac{1}{n}\frac{1}{n}\left[\sum_{k=1}^{n}\left[\begin{array}{l}1+\left\{\begin{array}{l}\bar{x}_1\left[S^{11(2)}\left(x_{k1}-\bar{x}_1\right)+S^{12(2)}\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{1m(2)}\left(x_{km}-\bar{x}_m\right)\right] \\ +\bar{x}_2\left[S^{21(2)}\left(x_{k1}-\bar{x}_1\right)+S^{22(2)}\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{2m(2)}\left(x_{km}-\bar{x}_m\right)\right] \\ +\cdots \\ +\bar{x}_m\left[S^{m1(2)}\left(x_{k1}-\bar{x}_1\right)+S^{m2(2)}\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{mm(2)}\left(x_{km}-\bar{x}_m\right)\right]\end{array}\right\}^2 \\ -2\left\{\begin{array}{l}\bar{x}_1\left[S^{11(2)}\left(x_{k1}-\bar{x}_1\right)+S^{12(2)}\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{1m(2)}\left(x_{km}-\bar{x}_m\right)\right] \\ +\bar{x}_2\left[S^{21(2)}\left(x_{k1}-\bar{x}_1\right)+S^{22(2)}\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{2m(2)}\left(x_{km}-\bar{x}_m\right)\right] \\ +\cdots \\ +\bar{x}_m\left[S^{m1(2)}\left(x_{k1}-\bar{x}_1\right)+S^{m2(2)}\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{mm(2)}\left(x_{km}-\bar{x}_m\right)\right]\end{array}\right\}\end{array}\right]\right]s_e^{(2)}
\end{aligned}
$$

(220)

ここで、式の中の係数-2 以降の項は

$$\sum_{i=1}^{n}\left(x_{ip}-\bar{x}_p\right)=0 \tag{221}$$

により、消える。したがって、

$$
\begin{aligned}
&V\left[\hat{a}_0\right] \\
&= \frac{1}{n}\frac{1}{n}\left[\sum_{k=1}^{n}\left[1+\left\{\begin{array}{l}
\bar{x}_1\left[S^{11(2)}\left(x_{k1}-\bar{x}_1\right)+S^{12(2)}\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{1m(2)}\left(x_{km}-\bar{x}_m\right)\right] \\
+\bar{x}_2\left[S^{21(2)}\left(x_{k1}-\bar{x}_1\right)+S^{22(2)}\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{2m(2)}\left(x_{km}-\bar{x}_m\right)\right] \\
+\cdots \\
+\bar{x}_m\left[S^{m1(2)}\left(x_{k1}-\bar{x}_1\right)+S^{m2(2)}\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{mm(2)}\left(x_{km}-\bar{x}_m\right)\right]
\end{array}\right\}^2\right]\right]S_e^{(2)} \\
&= \frac{1}{n}\frac{1}{n}\left[\sum_{k=1}^{n}\left[\begin{array}{l}
1 \\
+\bar{x}_1{}^2\left[S^{11(2)}\left(x_{k1}-\bar{x}_1\right)+S^{12(2)}\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{1m(2)}\left(x_{km}-\bar{x}_m\right)\right]^2 \\
+\bar{x}_2{}^2\left[S^{21(2)}\left(x_{k1}-\bar{x}_1\right)+S^{22(2)}\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{2m(2)}\left(x_{km}-\bar{x}_m\right)\right]^2 \\
+\cdots \\
+\bar{x}_m{}^2\left[S^{m1(2)}\left(x_{k1}-\bar{x}_1\right)+S^{m2(2)}\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{mm(2)}\left(x_{km}-\bar{x}_m\right)\right]^2 \\
+2\bar{x}_1\bar{x}_2\left[S^{11(2)}\left(x_{k1}-\bar{x}_1\right)+S^{12(2)}\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{1m(2)}\left(x_{km}-\bar{x}_m\right)\right]\left[S^{21(2)}\left(x_{k1}-\bar{x}_1\right)+S^{22(2)}\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{2m(2)}\right. \\
+\cdots \\
+2\bar{x}_{m-1}\bar{x}_m\left[S^{(m-1)(2)}\left(x_{k1}-\bar{x}_1\right)+S^{(m-1)(2)}\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{(m-1)m(2)}\left(x_{km}-\bar{x}_m\right)\right]\left[S^{m1(2)}\left(x_{k1}-\bar{x}_1\right)+S^{m2(2)}\left(x_{k2}-\bar{x}_2\right)+\right.
\end{array}\right]\right] \\
&= \frac{1}{n}+\frac{1}{n}\frac{1}{n}\left[\sum_{i=1}^{n}\left[\begin{array}{l}
\bar{x}_1{}^2\left[S^{11(2)}\left(x_{i1}-\bar{x}_1\right)+S^{12(2)}\left(x_{i2}-\bar{x}_2\right)+\cdots+S^{1m(2)}\left(x_{im}-\bar{x}_m\right)\right]^2 \\
+\bar{x}_2{}^2\left[S^{21(2)}\left(x_{i1}-\bar{x}_1\right)+S^{22(2)}\left(x_{i2}-\bar{x}_2\right)+\cdots+S^{2m(2)}\left(x_{im}-\bar{x}_m\right)\right]^2 \\
+\cdots \\
+\bar{x}_m{}^2\left[S^{m1(2)}\left(x_{i1}-\bar{x}_1\right)+S^{m2(2)}\left(x_{i2}-\bar{x}_2\right)+\cdots+S^{mm(2)}\left(x_{im}-\bar{x}_m\right)\right]^2 \\
+2\bar{x}_1\bar{x}_2\left[S^{11(2)}\left(x_{i1}-\bar{x}_1\right)+S^{12(2)}\left(x_{i2}-\bar{x}_2\right)+\cdots+S^{1m(2)}\left(x_{im}-\bar{x}_m\right)\right]\left[S^{21(2)}\left(x_{i1}-\bar{x}_1\right)+S^{22(2)}\left(x_{i2}-\bar{x}_2\right)+\cdots+S^{2m(}\right. \\
+\cdots \\
+2\bar{x}_{m-1}\bar{x}_m\left[S^{(m-1)(2)}\left(x_{i1}-\bar{x}_1\right)+S^{(m-1)(2)}\left(x_{i2}-\bar{x}_2\right)+\cdots+S^{(m-1)m(2)}\left(x_{im}-\bar{x}_m\right)\right]\left[S^{m1(2)}\left(x_{i1}-\bar{x}_1\right)+S^{m2(2)}\left(x_{i2}-\bar{x}_2\right)\right.
\end{array}\right]\right] \\
&= \frac{1}{n}\left\{1+\begin{pmatrix}\bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_m\end{pmatrix}\begin{pmatrix}S^{11(2)} & S^{12(2)} & \cdots & S^{1m(2)} \\ S^{21(2)} & S^{22(2)} & \cdots & S^{2m(2)} \\ \vdots & \vdots & \ddots & \vdots \\ S^{m1(2)} & S^{m2(2)} & \cdots & S^{mm(2)}\end{pmatrix}\begin{pmatrix}\bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_m\end{pmatrix}\right\}S_e^{(2)}
\end{aligned}
$$

(222)

$\hat{a}_1$ と $\hat{a}_2$ 間の共分散の期待値は以下のように求まる

$$
\begin{aligned}
&Cov\left[\hat{a}_1,\hat{a}_2\right]\\
&=\frac{1}{n}\frac{1}{n}\left\{\sum_{k=1}^{n}\left[\sum_{p=1}^{m}S^{1p(2)}\left(x_{kp}-\bar{x}_p\right)\right]\left[\sum_{p'=1}^{m}S^{2p'(2)}\left(x_{kp'}-\bar{x}_{p'}\right)\right]\right\}V\left[e\right]\\
&=\frac{1}{n}\frac{1}{n}\left\{\sum_{k=1}^{n}\left[S^{11(2)}\left(x_{k1}-\bar{x}_1\right)+S^{12(2)}\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{1m(2)}\left(x_{km}-\bar{x}_m\right)\right]\left[S^{21(2)}\left(x_{k1}-\bar{x}_1\right)+S^{22(2)}\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{2m(2)}\left(x_{km}-\bar{x}_n\right.\right.\right.\\
&=\frac{1}{n}\frac{1}{n}\sum_{k=1}^{n}\begin{bmatrix}S^{11(2)}S^{21(2)}\left(x_{k1}-\bar{x}_1\right)^2+S^{11(2)}S^{22(2)}\left(x_{k1}-\bar{x}_1\right)\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{11(2)}S^{2m(2)}\left(x_{k1}-\bar{x}_1\right)\left(x_{km}-\bar{x}_m\right)\\+S^{12(2)}S^{21(2)}\left(x_{k2}-\bar{x}_2\right)\left(x_{k1}-\bar{x}_1\right)+S^{12(2)}S^{22(2)}\left(x_{k2}-\bar{x}_2\right)^2+\cdots+S^{12(2)}S^{2m(2)}\left(x_{k2}-\bar{x}_2\right)\left(x_{km}-\bar{x}_m\right)\\+\cdots\\+S^{1m(2)}S^{2m(2)}\left(x_{km}-\bar{x}_m\right)\left(x_{km}-\bar{x}_m\right)\end{bmatrix}s_e^{(2)}\\
&=\frac{1}{n}\begin{bmatrix}S^{11(2)}\left[S^{21(2)}S_{11}^{(2)}+S^{22(2)}S_{12}^{(2)}+\cdots+S^{2m(2)}S_{1m}^{(2)}\right]\\+S^{12(2)}\left[S^{21(2)}S_{21}^{(2)}+S^{22(2)}S_{22}^{(2)}+\cdots+S^{2m(2)}S_{2m}^{(2)}\right]\\+\cdots\\+S^{1m(2)}\left[S^{21(2)}S_{m1}^{(2)}+S^{22(2)}S_{m2}^{(2)}+\cdots+S^{2m(2)}S_{mm}^{(2)}\right]\end{bmatrix}s_e^{(2)}\\
&=\frac{1}{n}S^{12(2)}s_e^{(2)}
\end{aligned}
$$

(223)

同様に

$$Cov\left[\hat{a}_2,\hat{a}_1\right]=\frac{1}{n}S^{21(2)}s_e^{(2)} \tag{224}$$

となる。一般に $\hat{a}_i$ と $\hat{a}_j$ $(i,j\neq 0)$ 間の共分散は

$$Cov\left[\hat{a}_j,\hat{a}_l\right]=\frac{1}{n}S^{jl(2)}s_e^{(2)} \tag{225}$$

となる。

$\hat{a}_0$ と $\hat{a}_1$ 間の共分散の期待値は

$$
\begin{aligned}
Cov\left[\hat{a}_0,\hat{a}_1\right] &= \frac{1}{n}\frac{1}{n}\sum_{k=1}^{n}\left\{1-\sum_{j=1}^{m}\bar{x}_j\sum_{p=1}^{m}S^{jp(2)}\left(x_{kp}-\bar{x}_p\right)\right\}\left[\sum_{p=1}^{m}S^{1p(2)}\left(x_{kp}-\bar{x}_p\right)\right]V[e] \\
&= \frac{1}{n}\frac{1}{n}\sum_{k=1}^{n}\left\{1-\sum_{j=1}^{m}\bar{x}_j\left[S^{j1(2)}\left(x_{k1}-\bar{x}_1\right)+S^{j2(2)}\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{jm(2)}\left(x_{km}-\bar{x}_m\right)\right]\right\}\left[S^{11(2)}\left(x_{k1}-\bar{x}_1\right)+S^{12(2)}\left(x_{k2}-\bar{x}_2\right)\right. \\
&= \frac{1}{n}\frac{1}{n}\sum_{k=1}^{n}\left\{\begin{array}{l}1\\ -\bar{x}_1\left[S^{11(2)}\left(x_{k1}-\bar{x}_1\right)+S^{12(2)}\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{1m(2)}\left(x_{km}-\bar{x}_m\right)\right]\\ -\bar{x}_2\left[S^{21(2)}\left(x_{k1}-\bar{x}_1\right)+S^{22(2)}\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{2m(2)}\left(x_{km}-\bar{x}_m\right)\right]\\ -\cdots\\ -\bar{x}_m\left[S^{m1(2)}\left(x_{k1}-\bar{x}_1\right)+S^{m2(2)}\left(x_{k2}-\bar{x}_2\right)+\cdots+S^{mm(2)}\left(x_{km}-\bar{x}_m\right)\right]\end{array}\right\}\left[S^{11(2)}\left(x_{k1}-\bar{x}_1\right)+S^{12(2)}\left(x_{k2}-\bar{x}_2\right)+\right. \\
&= \frac{1}{n}\frac{1}{n}\sum_{i=1}^{n}\left\{\begin{array}{l}1\\ -\bar{x}_1\left[S^{j1(2)}\left(x_{i1}-\bar{x}_1\right)+S^{j2(2)}\left(x_{i2}-\bar{x}_2\right)+\cdots+S^{m2(2)}\left(x_{im}-\bar{x}_m\right)\right]\\ -\bar{x}_2\left[S^{j1(2)}\left(x_{i1}-\bar{x}_1\right)+S^{j2(2)}\left(x_{i2}-\bar{x}_2\right)+\cdots+S^{m2(2)}\left(x_{im}-\bar{x}_m\right)\right]\\ -\cdots\\ -\bar{x}_m\left[S^{j1(2)}\left(x_{i1}-\bar{x}_1\right)+S^{j2(2)}\left(x_{i2}-\bar{x}_2\right)+\cdots+S^{m2(2)}\left(x_{im}-\bar{x}_m\right)\right]\end{array}\right\}\left[S^{11(2)}\left(x_{i1}-\bar{x}_1\right)+S^{12(2)}\left(x_{i2}-\bar{x}_2\right)+\cdots\right. \\
&= -\frac{1}{n}\left\{\bar{x}_1S^{11(2)}+\bar{x}_2S^{12(2)}+\cdots+\bar{x}_mS^{1m(2)}\right\}s_e^{(2)} \\
&= -\left(S^{11(2)}\quad S^{12(2)}\quad\cdots\quad S^{1m(2)}\right)\begin{pmatrix}\bar{x}_1\\ \bar{x}_2\\ \vdots\\ \bar{x}_m\end{pmatrix}s_e^{(2)}
\end{aligned}
\tag{226}
$$

となる。

一般に $Cov\left[\hat{a}_0,\hat{a}_j\right]$ は、

$$
Cov\left[\hat{a}_0,\hat{a}_j\right] = -\frac{1}{n}\sum_{l=1}^{m}\bar{x}_lS^{jl(2)}s_e^{(2)} \tag{227}
$$

となる。

係数 $\hat{a}_0,\hat{a}_1,\hat{a}_2,\cdots,\hat{a}_m$ は正規分布に従うとすると、規格化値

$$
t=\frac{\hat{a}_i-a_i}{\sqrt{\dfrac{S^{ii(2)}}{n}s_e^{(2)}}}\quad for\ i=0,1,2,\cdots,m \tag{228}
$$

は自由度ϕ_eのt分布に従う。ここで、自由度は

$$\phi_e = n-(m+1) \tag{229}$$

である。したがって、各係数の推定範囲は

$$\hat{a}_i - t_p(\phi_e;P)\sqrt{\frac{S^{ii(2)}}{n}s_e^{(2)}} \le a_i \le \hat{a}_i + t_p(\phi_e;P)\sqrt{\frac{S^{ii(2)}}{n}s_e^{(2)}} \tag{230}$$

となる。

4.3.8. 回帰値および目的変数の推定範囲

我々は、n セットのデータ$(x_{k1},x_{k2},\cdots,x_{km},y_k)$を得たとする。すると、対応する回帰値は、以下となる。

$$Y = \hat{a}_0 + \hat{a}_1x_1 + \hat{a}_2x_2 + \cdots + \hat{a}_mx_m \tag{231}$$

この期待値は

$$\begin{aligned} E[Y] &= E[\hat{a}_0] + E[\hat{a}_1]x_1 + E[\hat{a}_2]x_2 + \cdots + E[\hat{a}_m]x_m \\ &= a_0 + a_1x_1 + a_2x_2 + \cdots + a_mx_m \end{aligned} \tag{232}$$

この分散は

$$\begin{aligned} V[Y] &= V[\hat{a}_0 + \hat{a}_1x_1 + \hat{a}_2x_2 + \cdots + \hat{a}_mx_m] \\ &= \frac{1}{n}\left[1 + \begin{pmatrix} x_1-\bar{x}_1 & x_2-\bar{x}_2 & \cdots & x_m-\bar{x}_m \end{pmatrix} \begin{pmatrix} S^{11(2)} & S^{12(2)} & \cdots & S^{1m(2)} \\ S^{21(2)} & S^{22(2)} & \cdots & S^{2m(2)} \\ \cdots & \cdots & \ddots & \cdots \\ S^{m1(2)} & S^{m2(2)} & \cdots & S^{mm(2)} \end{pmatrix} \begin{pmatrix} x_1-\bar{x}_1 \\ x_2-\bar{x}_2 \\ \cdots \\ x_m-\bar{x}_m \end{pmatrix}\right] s_e^{(2)} \\ &= \frac{1}{n}\left(1+D^2\right)s_e^{(2)} \end{aligned} \tag{233}$$

となる。ただし、

$$D^2 = \begin{pmatrix} x_1-\bar{x}_1 & x_2-\bar{x}_2 & \cdots & x_m-\bar{x}_m \end{pmatrix} \begin{pmatrix} S^{11(2)} & S^{12(2)} & \cdots & S^{1m(2)} \\ S^{21(2)} & S^{22(2)} & \cdots & S^{2m(2)} \\ \cdots & \cdots & \ddots & \cdots \\ S^{m1(2)} & S^{m2(2)} & \cdots & S^{mm(2)} \end{pmatrix} \begin{pmatrix} x_1-\bar{x}_1 \\ x_2-\bar{x}_2 \\ \cdots \\ x_m-\bar{x}_m \end{pmatrix} \tag{234}$$

である。したがって、回帰値の推定値は以下のようになる。

$$\begin{aligned} &\hat{a}_0 + \hat{a}_1x_1 + \hat{a}_2x_2 + \cdots + \hat{a}_mx_m - t_p(\phi_e;P)\sqrt{\left(\frac{1}{n}+\frac{D^2}{n}\right)s_e^{(2)}} \\ &\le Y \le \\ &\hat{a}_0 + \hat{a}_1x_1 + \hat{a}_2x_2 + \cdots + \hat{a}_mx_m + t_p(\phi_e;P)\sqrt{\left(\frac{1}{n}+\frac{D^2}{n}\right)s_e^{(2)}} \end{aligned} \tag{235}$$

目的変数は、以下のようにあらわされる。

$$y = \hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2 + \cdots + \hat{a}_m x_m + e \tag{236}$$

したがって、対応する推定範囲は以下のようになる。

$$\begin{aligned} &\hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2 + \cdots + \hat{a}_m x_m - t(\phi_e, P)\sqrt{\left(1 + \frac{1}{n} + \frac{D^2}{n}\right) s_e^{(2)}} \\ &\leq y \leq \\ &\hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2 + \cdots + \hat{a}_m x_m + t(\phi_e, P)\sqrt{\left(1 + \frac{1}{n} + \frac{D^2}{n}\right) s_e^{(2)}} \end{aligned} \tag{237}$$

もしも、新たなデータがきて、それが上で評価される推定範囲外であれば、そのデータはこの集合に属していない、とみなされる。範囲内であれば、この集合に属している、とみなされる。

4.3.9. 残差誤差

決定係数、重相関係数、自由度調整済決定係数、F 値はデータ全体を評価するものであった。ここで、個別のデータがいいのかを判断する方法を示す。

残差誤差は

$$e_k = y_k - Y_k \tag{238}$$

で評価される。これは規格化して

$$z_{ek} = \frac{e_k}{\sqrt{s_e^{(2)}}} \tag{239}$$

となる。ここで、$s_e^{(2)}$ は

$$s_e^{(2)} = \frac{n}{\phi_e} S_e^{(2)} \tag{240}$$

で与えられる。この z_{ek} は標準正規分布 $N(0,1^2)$ に従う。したがって、これから、この誤差がある有意値 z_P より大きいかどうかで、データを判断できる。このデータ評価では、どの位置のデータなのかの情報を利用していない。つまり、データ位置に関わりなくその誤差を評価している。

4.3.10. 梃子比

位置によるデータの誤差を考慮するため、別な視点から評価する。

k 番目の回帰値は以下のように与えられる。

$$
\begin{aligned}
Y_k &= \hat{a}_0 + \hat{a}_1 x_{k1} + \hat{a}_2 x_{k2} + \cdots + \hat{a}_m x_{km} \\
&= \bar{y} - \hat{a}_1 \left(x_{k1} - \bar{x}_1 \right) + \hat{a}_2 \left(x_{k2} - \bar{x}_2 \right) + \cdots + \hat{a}_m \left(x_{km} - \bar{x}_m \right) \\
&= \bar{y} - \left(x_{k1} - \bar{x}_1 \quad x_{k2} - \bar{x}_2 \quad \cdots \quad x_{km} - \bar{x}_m \right) \begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_m \end{pmatrix}
\end{aligned} \tag{241}
$$

右辺の項はそれぞれ以下のように展開される。

$$
\bar{y} = \frac{1}{n} \sum_{k=1}^{n} y_k \tag{242}
$$

$$
\begin{aligned}
&\left(x_{k1} - \bar{x}_1 \quad x_{k2} - \bar{x}_2 \quad \cdots \quad x_{km} - \bar{x}_m \right) \begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_m \end{pmatrix} \\
&= \left(x_{k1} - \bar{x}_1 \quad x_{k2} - \bar{x}_2 \quad \cdots \quad x_{km} - \bar{x}_m \right) \begin{pmatrix} S_{11}^{(2)} & S_{12}^{(2)} & \cdots & S_{1m}^{(2)} \\ S_{21}^{(2)} & S_{22}^{(2)} & \cdots & S_{2m}^{(2)} \\ \cdots & \cdots & \ddots & \cdots \\ S_{m1}^{(2)} & S_{m2}^{(2)} & \cdots & S_{mm}^{(2)} \end{pmatrix}^{-1} \begin{pmatrix} S_{1y}^{(2)} \\ S_{2y}^{(2)} \\ \cdots \\ S_{my}^{(2)} \end{pmatrix} \\
&= \left(x_{k1} - \bar{x}_1 \quad x_{k2} - \bar{x}_2 \quad \cdots \quad x_{km} - \bar{x}_m \right) \begin{pmatrix} S^{11(2)} & S^{12(2)} & \cdots & S^{1m(2)} \\ S^{21(2)} & S^{22(2)} & \cdots & S^{2m(2)} \\ \cdots & \cdots & \ddots & \cdots \\ S^{m1(2)} & S^{m2(2)} & \cdots & S^{mm(2)} \end{pmatrix} \begin{pmatrix} S_{1y}^{(2)} \\ S_{2y}^{(2)} \\ \cdots \\ S_{my}^{(2)} \end{pmatrix} \\
&= \left(x_{k1} - \bar{x}_1 \quad x_{k2} - \bar{x}_2 \quad \cdots \quad x_{km} - \bar{x}_m \right) \begin{pmatrix} S^{11(2)} & S^{12(2)} & \cdots & S^{1m(2)} \\ S^{21(2)} & S^{22(2)} & \cdots & S^{2m(2)} \\ \cdots & \cdots & \ddots & \cdots \\ S^{m1(2)} & S^{m2(2)} & \cdots & S^{mm(2)} \end{pmatrix} \begin{pmatrix} \frac{1}{n} \sum_{k=1}^{n} \left(x_{k1} - \bar{x}_1 \right) \left(y_k - \bar{y} \right) \\ \frac{1}{n} \sum_{k=1}^{n} \left(x_{k2} - \bar{x}_2 \right) \left(y_k - \bar{y} \right) \\ \cdots \\ \frac{1}{n} \sum_{k=1}^{n} \left(x_{km} - \bar{x}_m \right) \left(y_k - \bar{y} \right) \end{pmatrix} \\
&= \left(x_{k1} - \bar{x}_1 \quad x_{k2} - \bar{x}_2 \quad \cdots \quad x_{km} - \bar{x}_m \right) \begin{pmatrix} S^{11(2)} & S^{12(2)} & \cdots & S^{1m(2)} \\ S^{21(2)} & S^{22(2)} & \cdots & S^{2m(2)} \\ \cdots & \cdots & \ddots & \cdots \\ S^{m1(2)} & S^{m2(2)} & \cdots & S^{mm(2)} \end{pmatrix} \begin{pmatrix} \frac{1}{n} \sum_{k=1}^{n} \left(x_{k1} - \bar{x}_1 \right) y_k \\ \frac{1}{n} \sum_{k=1}^{n} \left(x_{k2} - \bar{x}_2 \right) y_k \\ \cdots \\ \frac{1}{n} \sum_{k=1}^{n} \left(x_{km} - \bar{x}_m \right) y_k \end{pmatrix}
\end{aligned} \tag{243}
$$

したがって、回帰値は

$$Y_k = h_{k1}y_1 + h_{k2}y_2 + \cdots + h_{kk}y_k + \cdots + h_{kn}y_n \tag{244}$$

と表現される。ここで、

$$h_{kl} = \frac{1}{n} + \frac{1}{n}\begin{pmatrix} x_{k1} - \overline{x}_1 & x_{k2} - \overline{x}_2 & \cdots & x_{km} - \overline{x}_m \end{pmatrix} \begin{pmatrix} S^{11(2)} & S^{12(2)} & \cdots & S^{1m(2)} \\ S^{21(2)} & S^{22(2)} & \cdots & S^{2m(2)} \\ \cdots & \cdots & \ddots & \cdots \\ S^{m1(2)} & S^{m2(2)} & \cdots & S^{mm(2)} \end{pmatrix} \begin{pmatrix} x_{l1} - \overline{x}_1 \\ x_{l2} - \overline{x}_2 \\ \vdots \\ x_{lm} - \overline{x}_m \end{pmatrix} \tag{245}$$

である。ここで、k 番目の係数すなわち $h_{k,l=k} = h_{kk}$ に注目する。それは

$$h_{kk} = \frac{1}{n} + \frac{D_k^2}{n} \tag{246}$$

である。ただし、

$$D_k^2 = \begin{pmatrix} x_{k1} - \overline{x}_1 & x_{k2} - \overline{x}_2 & \cdots & x_{km} - \overline{x}_m \end{pmatrix} \begin{pmatrix} S^{11(2)} & S^{12(2)} & \cdots & S^{1m(2)} \\ S^{21(2)} & S^{22(2)} & \cdots & S^{2m(2)} \\ \cdots & \cdots & \ddots & \cdots \\ S^{m1(2)} & S^{m2(2)} & \cdots & S^{mm(2)} \end{pmatrix} \begin{pmatrix} x_{k1} - \overline{x}_1 \\ x_{k2} - \overline{x}_2 \\ \cdots \\ x_{km} - \overline{x}_m \end{pmatrix} \tag{247}$$

D_k^2 はマハラビノスの距離の平方と呼ばれる。

挺子比の和は

$$\begin{aligned} \sum_{k=1}^{n} h_{kk} &= \sum_{k=1}^{n} \frac{1}{n} + \sum_{k=1}^{n} \frac{D_k^2}{n} \\ &= 1 + \frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{m} \sum_{j=1}^{m} \left(x_{ki} - \overline{x}_i \right) \left(x_{kj} - \overline{x}_j \right) S^{ij(2)} \\ &= 1 + \sum_{i=1}^{m} \sum_{i=1}^{m} \left[\frac{\sum_{k=1}^{n} \left(x_{ki} - \overline{x}_i \right) \left(x_{kj} - \overline{x}_j \right)}{n} \right] S^{ij(2)} \\ &= 1 + \sum_{i=1}^{m} \sum_{i=1}^{m} S_{ij}^{(2)} S^{ij(2)} \end{aligned} \tag{248}$$

である。ここで、

$$\left(S_{ij}^{(2)} \right) \left(S^{ij(2)} \right) = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix} \tag{249}$$

であるから、挺子比の和は

$$\begin{aligned} \sum_{k=1}^{n} h_{kk} &= 1 + \sum_{i=1}^{m} \sum_{j=1}^{m} S_{ij}^{(2)} S^{ij(2)} \\ &= 1 + m \end{aligned} \tag{250}$$

となる。ここで、

$$S_{ij}^{(2)} S^{ij(2)} = \delta_{ij} \tag{251}$$

であり、$i = j$ となるのは $i = 1,2,\cdots,m$ の m 個であることを利用している。

したがって、梃子比の平均は

$$\mu_{hkk} = \frac{m+1}{n} \tag{252}$$

したがって、梃子比 h_{kk} の臨界値として

$$h_{kk} \le \kappa \times \mu_{hkk} \tag{253}$$

が採用される場合がある。$\kappa = 2$ が良く用いられる。梃子比が $\kappa\mu_{h_{kk}}$ よりも大きい場合は外れ値として扱われる場合がある。

4.3.11. 誤差分散と梃子比を考慮した誤差

以上より、誤差分散と梃子比を組み合わせて誤差の評価をするのが尤もらしい。それは、

$$t = \frac{z_{ek}}{\sqrt{1 - h_{kk}}} \tag{254}$$

で与えられる。我々は、t に対して臨界値を与え、それを超えていれば外れ値として扱うことができる。

4.3.12. 重回帰係数と相関係数の関係

ここでは、重回帰係数と相関係数の関係を議論する。ここでは、二つの説明変数 x_1 と x_2 の場合を扱い、かつそれぞれは規格化されているものとする。

重回帰係数と相関係数は以下のように関連づけられる。

$$\begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} \tag{255}$$

したがって、それぞれの係数は以下のようになる。

$$\begin{aligned} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} &= \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}^{-1} \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} \\ &= \begin{pmatrix} \dfrac{1}{1-{r_{12}}^2} & -\dfrac{r_{12}}{1-{r_{12}}^2} \\ -\dfrac{r_{12}}{1-{r_{12}}^2} & \dfrac{1}{1-{r_{12}}^2} \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} \\ &= \begin{pmatrix} \dfrac{r_1 - r_{12} r_2}{1-{r_{12}}^2} \\ \dfrac{r_2 - r_{12} r_1}{1-{r_{12}}^2} \end{pmatrix} \end{aligned} \tag{256}$$

したがって、以下のようになる。

$$a_1 = r_1 \frac{1 - r_{12}\dfrac{r_2}{r_1}}{1 - {r_{12}}^2} \tag{257}$$

$$a_2 = r_2 \frac{1 - r_{12}\dfrac{r_1}{r_2}}{1 - {r_{12}}^2} \tag{258}$$

ここで $r_1 \geq r_2$ を仮定する。

もし二つの変数が独立であれば、つまり $r_{12} = 0$ であれば、

$$a_1 = r_1 \tag{259}$$

$$a_2 = r_2 \tag{260}$$

となる。そして、それぞれの係数は対応する変数と目的変数との相関係数に一致する。

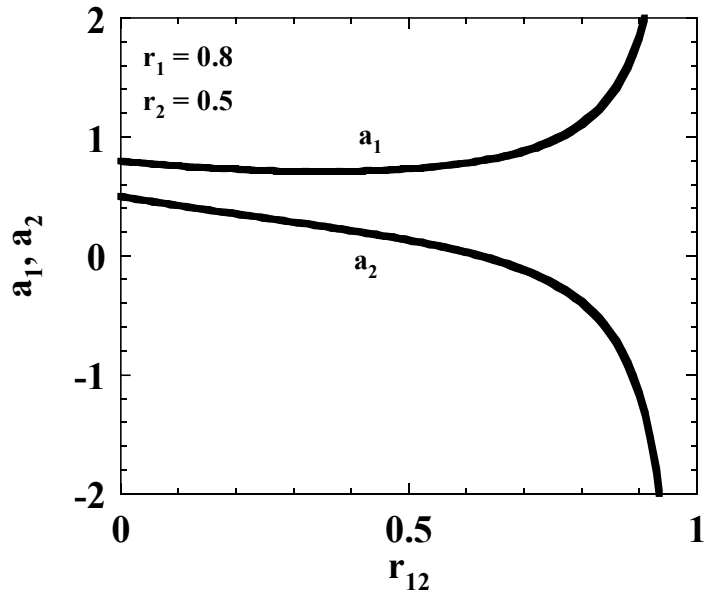

Figure 1 重相関係数と相関係数の関係.

Figure 1 は変数 1,2 間の相関係数と重回帰係数の関連を示している。係数 a_1 は、r_{12} が小さい場合は、それが増加するにつれわずかに低下していくが、その後急激に増加する。一方、係数 a_2 は、r_{12} が小さい場合は、それが増加するにつれわずかに低下していくが、その後急激に減少する。つまり、説明変数間の相互作用が小さい場合は重回帰係数と各変数と目的変数との相関係数間の値はそれほど大きくずれないが、説明変数間の相関係数が大きくなると、重回帰係数はそれぞれの相関係数の値とはかけ離れてしまう。

Figure 1 の依存性を解析する。

a_1 を r_{12} で偏微分して

$$\frac{\partial a_1}{\partial r_{12}} = r_1 \frac{-\frac{r_2}{r_1}\left(1-{r_{12}}^2\right)-\left(1-r_{12}\frac{r_2}{r_1}\right)(-2r_{12})}{\left(1-{r_{12}}^2\right)^2} = \frac{-r_2}{\left(1-{r_{12}}^2\right)^2}\left({r_{12}}^2-2\frac{r_1}{r_2}r_{12}+1\right) \tag{261}$$

となる。これを 0 と置き、以下を得る。

$${r_{12}}^2-2\frac{r_1}{r_2}r_{12}+1=0 \tag{262}$$

これを解くと

$$r_{12}=\frac{r_1}{r_2}\pm\sqrt{\left(\frac{r_1}{r_2}\right)^2-1} \tag{263}$$

となる。ここで、r_{12} は 1 より小さいから、上の解の符号の正のものは解として不適である。

したがって、解は以下のようになる。

$$r_{12}=\frac{r_1}{r_2}-\sqrt{\left(\frac{r_1}{r_2}\right)^2-1} = \frac{1}{\frac{r_1}{r_2}+\sqrt{\left(\frac{r_1}{r_2}\right)^2-1}} \tag{264}$$

この r_{12} において、a_1 は最小になる。この点が、ほぼ重回帰係数と相関係数の定性的な関連性が維持されているとみなすことができる。これいじょうだと、重回帰係数から各説明変数と目的変数の関連性を予想することは難しくなる。

4.3.13. 偏相関係数

二変数の場合の擬似相関、偏相関係数を議論した。その解析は多変量の場合に拡張される。

これまで扱ってきたように目的変数 y と m 個の説明変数 $x_1, x_2, \cdots, x_m$ を考える。

この中で y と x_1 の相関を考える。この場合、y も x_1 も一般には残りの説明変数 $x_2, \cdots, x_m$ と相関を持っている。ここで知りたいのは、その相関を除いた純粋の y と x_1 の相関である。それを評価するために以下のように解析する。

x_1 のデータを除いたものから y の回帰式

$$Y_k = c_0 + c_2 x_{k2} + c_3 x_{k3} + \cdots + c_m x_{km} \tag{265}$$

を得ることができる。ここで、

$$\begin{pmatrix} c_2 \\ c_3 \\ \cdots \\ c_m \end{pmatrix} = \begin{pmatrix} S_{22}^{(2)} & S_{23}^{(2)} & \cdots & S_{2m}^{(2)} \\ S_{32}^{(2)} & S_{33}^{(2)} & \cdots & S_{3m}^{(2)} \\ \cdots & \cdots & \ddots & \cdots \\ S_{m2}^{(2)} & S_{m3}^{(2)} & \cdots & S_{mm}^{(2)} \end{pmatrix}^{-1} \begin{pmatrix} S_{2y}^{(2)} \\ S_{3y}^{(2)} \\ \cdots \\ S_{my}^{(2)} \end{pmatrix} \tag{266}$$

となる。これから $c_2, c_3, \cdots, c_m$ が求まる。$c_2, c_3, \cdots, c_m$ から c_0 は以下のように求まる。

$$\bar{y} = c_0 + c_2 \bar{x}_2 + c_3 \bar{x}_3 + \cdots + c_m \bar{x}_m \tag{267}$$

一方、y を除き、x_1 を目的変数とみなすとその回帰式は

$$X_{k1} = d_0 + d_2 x_{k2} + d_3 x_{k3} + \cdots + d_m x_{km} \tag{268}$$

となる。ここで、各係数 $d_2, d_3, \cdots, d_m$ は

$$\begin{pmatrix} d_2 \\ d_3 \\ \cdots \\ d_p \end{pmatrix} = \begin{pmatrix} S_{22}^{(2)} & S_{23}^{(2)} & \cdots & S_{2m}^{(2)} \\ S_{32}^{(2)} & S_{33}^{(2)} & \cdots & S_{3m}^{(2)} \\ \cdots & \cdots & \ddots & \cdots \\ S_{m2}^{(2)} & S_{m3}^{(2)} & \cdots & S_{mm}^{(2)} \end{pmatrix}^{-1} \begin{pmatrix} S_{21}^{(2)} \\ S_{31}^{(2)} \\ \cdots \\ S_{m1}^{(2)} \end{pmatrix} \tag{269}$$

から求まる。d_0 は以下から求まる。

$$\bar{x}_1 = d_0 + d_2 \bar{x}_2 + d_3 \bar{x}_3 + \cdots + d_m \bar{x}_m \tag{270}$$

ここで、以下の変数を導入する。

$$u_k = y_k - Y_k \tag{271}$$

$$v_{k1} = x_{k1} - X_{k1} \tag{272}$$

これらの変数の平均と分散は以下のようになる。

$$\bar{u} = \frac{1}{n} \sum_{k=1}^{n} u_k \tag{273}$$

$$\bar{v}_1 = \frac{1}{n} \sum_{k=1}^{n} v_{k1} \tag{274}$$

$$S_{uu}^{(2)} = \frac{1}{n} \sum_{k=1}^{n} \left(u_k - \bar{u} \right)^2 \tag{275}$$

$$S_{v_1 v_1}^{(2)} = \frac{1}{n} \sum_{k=1}^{n} \left(v_{k1} - \bar{v}_1 \right)^2 \tag{276}$$

$$S_{uv_1}^{(2)} = \frac{1}{n}\sum_{k=1}^{n}\left(u_k - \bar{u}\right)\left(v_{k1} - \bar{v}_1\right) \tag{277}$$

したがって、他の変数の影響を除いた y と x_1 の相関、不偏相関 $r_{(1y),(2,3,\cdots,m)}$ は

$$r_{(1y),(2,3,\cdots,m)} = \frac{S_{uv_1}^{(2)}}{\sqrt{S_{v_1v_1}^{(2)}}\sqrt{S_{uu}^{(2)}}} \tag{278}$$

となる。他の変数 $x_2, x_3, \cdots, x_m$ についても同様に求めることができる。

4.3.14. 多重共線性

ある説明変数が他の説明変数の 1 次式で近似的に表される場合は、回帰係数の推定値の誤差は大きくなる。このように、説明変数間に強い相関がある場合に起こることを多重共線性と呼ぶ。

ドルベース輸出価格指数 y を工業製品卸売物価指数 x_1、先進国工業製品輸出価格指数 x_2 で表すことを考える。この重回帰式を求めると

$$y = 11.010 - 0.019x_1 + 0.933x_2 \tag{279}$$

となる。これらの相関係数は

Table 1 輸出価格指数 y を工業製品卸売物価指数 x_1、先進国工業製品輸出価格指数 x_2 間の相関係数

	x_1	x_2	y
x_1	1		
x_2	0.900	1	
y	0.871	0.990	1

この場合は説明変数間の酢関係数は 0.900 と大きい。推定した重回帰式の誤差が大きくなる。

それぞれの変数との回帰直線を求めると

$$y = -33.3 + 1.42x_1 \tag{280}$$

$$y = -10.2 + 0.923x_2 \tag{281}$$

となる。つまり、それぞれの回帰係数は正となる。しかしながら、y の x_1, x_2 に対する重回帰式においては、$x_1,$ に対するけ係数は負になっている。これは工業製品卸売物価指数が高くなればなるほど輸出価格指数は安くなるというもので、定性的な感覚とずれている。

これは、それぞれの係数の分散が

となり、二変数の相関が大きい場合はその分散が大きくなることに起因する。

ここでは、説明変数を二つとして解析したが、多くの説明変数の中にいくつかの説明変数間の相関係数がおおきくなれば、その係数に対する分散がおおきくなることを示される。

4.3.15. 変数の選択

変数の数を増やせば、一般にデータと回帰式の一致度はよくなる。しかし、目的変数に対して効いている説明変数のみを扱いたい、という要望がある。この説明変数の選択をするプロセスを検討する。

まず、目的変数に対して各説明変数の相関係数を評価する。そして、その中で最大の絶対値を持つ相関係数の項目を選択する。

その相関の有無を検定する。検定の結果が相関ありとみなせるならば、次のプロセスにすすむ。相関なしとみなせるならば、解析は終了する。

その説明変数を用いて回帰分析を行う。その係数が 0 とみなすせるかの検定を行う。

検定の結果 0 とみなせないならば、次のプロセスへすすむ。0 とみなせるならば解析は終了する。

次に、第 1 番目の変数の影響を除いた偏相関係数を評価する。そして、その中で最大の絶対値を持つ相関係数の項目を選択する。

その相関の有無を検定する。検定の結果が相関ありとみなせるならば、次のプロセスにすすむ。相関なしとみなせるならば、解析は終了する。

これまで選択した説明変数を用いて回帰分析を行う。各係数が 0 とみなすせるかの検定を行う。

検定の結果 0 とみなせないならば、次のプロセスへすすむ。0 とみなせる係数があるならば、対応する項目をはずす。

次に、第 1 、2 番目の変数の影響を除いた偏相関係数を評価する。そして、その中で最大の絶対値を持つ相関係数の項目を選択する。

その相関の有無を検定する。検定の結果が相関ありとみなせるならば、次のプロセスにすすむ。相関なしとみなせるならば、解析は終了する。

これまで選択した説明変数を用いて回帰分析を行う。各係数が 0 とみなすせるかの検定を行う。

検定の結果 0 とみなせないならば、次のプロセスへすすむ。0 とみなせる係数があるならば、対応する項目をはずす。

全ての変数をチェックしたかをみる。しているならば、終了する。したいないならば、以上のプロセスを繰り返す。

4.3.16. 変数の選択 2

変数の数を増やせば、一般にデータと回帰式の一致度はよくなる。しかし、目的変数に対して効いている説明変数のみを扱いたい、という要望がある。この説明変数の選択をするプロセスを検討する。これを行うには、変数の有効性とは何かを定量的に定義しなければならない。前述のとおり、効いていない変数でも追加すれば、変数の自由度を増加させ、データとの一致度はよくなるから、単純に誤差分散の大小を評価しても変数の有効性は評価できない。

変数を追加したことによって変動した回帰に起因する分散が、変数追加によって現象した誤差分散より大きければ、その変数は有効である、と判断する。

【Method 1】

まず、目的変数 y のみを考える、それに対応するモデル番号を 0 とする。

y の平均と分散を以下のように求める。

平均は以下である。

$$Model0 : Y_i = \bar{y} \tag{282}$$

分散 $S^{(2)}_{e(M0)}$ は以下である。

$$S^{(2)}_{e(M0)} = \frac{1}{n}\sum_{k=1}^{n}\left(y_k - Y_k\right)^2 = \frac{1}{n}\sum_{k=1}^{n}\left(y_k - \bar{y}\right)^2 = S^{(2)}_{yy} \tag{283}$$

次のステップとして、各説明変数 x_1，x_2，および x_m，をそれぞれ考える。それぞれの説明変数に関して、回帰値を

$$Y_k = a_0 + a_1 x_{kl_1} \tag{284}$$

を定義する。これに対応する回帰分散は $S^{(2)}_{e(M1)}$ である。

$$S_{e(M1)}^{(2)} = \frac{1}{n}\sum_{k=1}^{n}\left(y_k - Y_k\right)^2 = \frac{1}{n}\sum_{k=1}^{n}\left[y_k - \left(a_0 + a_1 x_{kl_1}\right)\right]^2 \tag{285}$$

そして、以下の F1 値を評価する。

$$F_1 = \frac{\left(nS_{e(M0)}^{(2)} - nS_{e(M1)}^{(2)}\right)\Big/\left(\phi_{e(M0)} - \phi_{e(M1)}\right)}{nS_{e(M1)}^{(2)}\Big/\phi_{e(M1)}} \tag{286}$$

これは F 分布に従い、その自由ド $\phi_{e(M0)}$ および $\phi_{e(M1)}$ とは以下で与えられる。

$$\phi_{e(M0)} = n - 1 \tag{287}$$

$$\phi_{e(M1)} = n - 2 \tag{288}$$

したがって、以下の F_1 の評価をする。

$$\begin{cases} F_1 \geq F\left(\phi_{e(M0)} - \phi_{e(M1)}, \phi_{e(M1)}\right) & valid \\ F_1 < F\left(\phi_{e(M0)} - \phi_{e(M1)}, \phi_{e(M1)}\right) & invalid \end{cases} \tag{289}$$

もしも有効な変数がなければ以上のプロセスは終わる。

もしも有効な変数があれば、その最大値を与える変数を選択する。

次に、二番目の変数を評価する。それをモデル 2 とする。この場合の回帰値は以下で与えられる。

$$Y_i = a_0 + a_1 x_{il_1} + a_2 x_{il_2} \tag{290}$$

関連する分散は以下で与えられる。

$$S_{e(M2)}^{(2)} = \frac{1}{n}\sum_{k=1}^{n}\left(y_k - Y_k\right)^2 = \frac{1}{n}\sum_{k=1}^{n}\left[y_k - \left(a_0 + a_1 x_{kl_1} + a_2 x_{kl_2}\right)\right]^2 \tag{291}$$

この比からなる以下の変数 F_2 を考える。

$$F_2 = \frac{\left(nS_{e(M1)}^{(2)} - nS_{e(M2)}^{(2)}\right)\Big/\left(\phi_{e(M1)} - \phi_{e(M2)}\right)}{nS_{e(M2)}^{(2)}\Big/\phi_{e(M2)}} \tag{292}$$

ここで、この変数は F 分布に従う。ここで自由度 $\phi_{e(M1)}$ および $\phi_{e(M2)}$ は以下のようになる。

$$\phi_{e(M1)} = n - 2 \tag{293}$$

$$\phi_{e(M2)} = n - 3 \tag{294}$$

以下の判断をすればいい。

$$\begin{cases} F_2 \geq F\left(\phi_{e(M1)} - \phi_{e(M2)}, \phi_{e(M2)}\right) & valid \\ F_2 < F\left(\phi_{e(M1)} - \phi_{e(M2)}, \phi_{e(M2)}\right) & invalid \end{cases} \tag{295}$$

もしも有効な変数がなければ以上のプロセスは終わる。

もしも有効な変数があれば、その最大値を与える変数を選択する。

次に、3 番目の変数を選択する。

以上のプロセスを、有効な変数が無くなるか、全て種類の変数を評価するまで続ける。そして、有効な変数を同定する。

4.3.17. ロジスティック回帰

ロジステック分析は、回帰分析の章で紹介した。これは、そこで述べたように医学の疫学研究の分野で開発された分析方法である。

これは、目的変数が発症率のように 0 から 1 をとる場合に適用される。その目的変数と関連する説明変数は回帰分析の場合は 1 種類のみであった。しかし、その原因候補となる説明変数は 1 種類ではなく複数の場合がある。したがって、ロジスティック分析は重回帰に拡張する必要がある。

例えば、回帰分析の章であ使った心臓病について考えてみる。回帰分析の章では、心臓病の原因は血圧だけであると仮定している。しかし、年齢、血圧、心拍数などのデータを加えるとより精度が高くなると思われる。年齢が 40 歳で血圧が 130mmHg、心拍数 80bpm の人の発生率が 0.3 である、というようなデータが多く集められているとする。このとき、年齢、血圧、心拍数から発症率を予測するのがロジスティック分析である。

年齢、血圧、心拍数、発症率は量的データであるから、発症率を予測するには単純に回帰分析を利用すればいいように思う。しかし、通常の直線型の回帰分析では年齢、血圧、心拍数の値によっては発症率は負や 1 より大きい値になる。一方、発症率のとる値は 0 から 1 である。

そこでロジスティック分析では対応する関数

$$Y = \frac{ae^{\sum_{k=1}^{m} b_k x_k}}{1 + ae^{\sum_{k=1}^{m} b_k x_k}} = \frac{1}{1 + \dfrac{1}{ae^{\sum_{k=1}^{m} b_k x_k}}} \tag{296}$$

を定義して、それに対して回帰を行う。これは、式を変形すると

$$\frac{Y}{1-Y}=\frac{1}{1+\dfrac{1}{ae^{\sum_{k=1}^{m}b_k x_k}}}\frac{1}{\left(1-\dfrac{1}{1+\dfrac{1}{ae^{\sum_{k=1}^{m}b_k x_k}}}\right)}$$

$$=\frac{1}{1+\dfrac{1}{ae^{\sum_{k=1}^{m}b_k x_k}}}\frac{1}{\dfrac{ae^{\sum_{k=1}^{m}b_k x_k}}{1+\dfrac{1}{ae^{\sum_{k=1}^{m}b_k x_k}}}} \tag{297}$$

$$=ae^{\sum_{k=1}^{m}b_k x_k}$$

となる。この両辺の対数をとると

$$\ln\left(\frac{Y}{1-Y}\right)=\ln a+\sum_{k=1}^{m}b_k x_k \tag{298}$$

となる。

よって、各データは

$$\ln\left(\frac{y_k}{1-y_k}\right)=\ln a+\sum_{k=1}^{m}b_k x_k+e_k \tag{299}$$

と表現される。あとは、通常の重回帰分析を行えばいい。

4.4. 重回帰分析の変数に対するコメント

ここでは、重回帰分析において説明変数の数の定め方を示した。

ここでは説明変数がどうあるべきかを簡単に記述する。

重回帰分析では説明変数は複数あり、その説明変数間の相互作用も考慮していた。そのため精度は向上する。施策においてその説明変数をどうにかする場合と何もしない場合がある。

説明変数に対して何もしない場合は、この解析で問題ない。精度が高いほうがいい。

しかし、説明変数に対して何かを施して目的変数を向上させようとすると問題がある。この解析においては説明変数間の相互作用を考慮している。したがって、説明変数同士は独立ではない。ある説明変数をいじると、それが滅の変数にも影響を及ぼしてしまう。それがどの程度であるのか定量的に予想することは困難である。したがって、この場合は説明変数間の相互作用は無視できるようなものを選択する。そして、各説明変数は独立であるとして解析をするめていく。

説明変数間の相関係数は選択した説明変数が妥当であるかどうかのみに利用する。そして、説明変数間の相互作用なないとして分析していく。

4.5. 重回帰分析の行列表現

重回帰分析は行列表現できる。行列表現の利点はその表現形式が次元つまり説明変数の数によらないということである。

4.5.1. 回帰

以下の行列、ベクトルを定義する。

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \tag{300}$$

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \tag{301}$$

$$\mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} \tag{302}$$

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} \tag{303}$$

$$\Sigma = \begin{pmatrix} S_{11}^{(2)} & S_{12}^{(2)} & \cdots & S_{1m}^{(2)} \\ S_{21}^{(2)} & S_{22}^{(2)} & \cdots & S_{2m}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ S_{m1}^{(2)} & S_{m2}^{(2)} & \cdots & S_{mm}^{(2)} \end{pmatrix} \tag{304}$$

$$\mathbf{S}_y^{(2)} = \begin{pmatrix} S_{1y}^{(2)} \\ S_{2y}^{(2)} \\ \vdots \\ S_{my}^{(2)} \end{pmatrix} \tag{305}$$

$$\boldsymbol{\alpha} = \begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_m \end{pmatrix} \tag{306}$$

$$\boldsymbol{\mu}_x = \begin{pmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_m \end{pmatrix} \tag{307}$$

新たなデータ

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} \tag{308}$$

既存のk番目のデータ

$$\mathbf{x}_k = \begin{pmatrix} x_{k1} \\ x_{k2} \\ \vdots \\ x_{km} \end{pmatrix} \tag{309}$$

$$\overline{x}_i = \frac{1}{n}\sum_{k=1}^{n} x_{ki} \tag{310}$$

$$\overline{y} = \frac{1}{n}\sum_{k=1}^{n} y_k \tag{311}$$

$$S_{ij}^{(2)} = \frac{1}{n}\sum_{k=1}^{n}\left(x_{ki} - \overline{x}_i\right)\left(x_{kj} - \overline{x}_j\right) \tag{312}$$

$$S_{iy}^{(2)} = \frac{1}{n}\sum_{k=1}^{n}\left(x_{ki} - \overline{x}_i\right)\left(y_k - \overline{y}\right) \tag{313}$$

$$S_{yy}^{(2)} = \frac{1}{n}\sum_{k=1}^{n}\left(y_k - \overline{y}\right)^2 \tag{314}$$

$$S_e^{(2)} = \frac{1}{n}\sum_{k=1}^{n}\left(y_k - Y_k\right)^2 = \frac{1}{n}\mathbf{e}^T\mathbf{e} \tag{315}$$

$$S_R^{(2)} = \frac{1}{n}\sum_{k=1}^{n}\left(Y_k - \bar{y}\right)^2 \tag{316}$$

$$\boldsymbol{\alpha} = \Sigma^{-1}\mathbf{S}_y^{(2)} \tag{317}$$

$$\hat{a}_0 = \bar{y} - \boldsymbol{\alpha}^T\boldsymbol{\mu}_x \tag{318}$$

4.5.2. 回帰の精度

$$R^2 = \frac{S_R^{(2)}}{S_{yy}^{(2)}} \tag{319}$$

$$S_R^{(2)} = \boldsymbol{\alpha}^T\mathbf{S}_y^{(2)} \tag{320}$$

$$R^{*2} = 1 - \frac{\frac{n}{\phi_e}S_e^{(2)}}{\frac{n}{\phi_T}S_{yy}^{(2)}} \tag{321}$$

$$\phi_e = n - (m+1) \tag{322}$$

である。

各分散の関係は

$$S_{yy}^{(2)} = S_R^{(2)} + S_e^{(2)} \tag{323}$$

であり、これは自由度の関係の方程式と直接関連する。すなわち

$$\phi_T = \phi_R + \phi_e \tag{324}$$

$$F = \frac{s_r^{(2)}}{s_e^{(2)}} \tag{325}$$

が大きければ、回帰の精度はいいといえる。ただし、

$$s_R^{(2)} = \frac{nS_R^{(2)}}{m} \tag{326}$$

$$s_e^{(2)} = \frac{nS_e^{(2)}}{n-(m+1)} \tag{327}$$

$$\begin{cases} F \le F_p\left(m, n-(m+1)\right) \Rightarrow invalid \\ F > F_p\left(m, n-(m+1)\right) \Rightarrow valid \end{cases} \tag{328}$$

4.5.3. データの精度

$$z_{ek} = \frac{e_k}{\sqrt{S_e^{(2)}}} \tag{329}$$

$$h_{kk} = \frac{1}{n} + \frac{D_k^2}{n} \tag{330}$$

$$D_k^2 = \left(\mathbf{x}_k - \boldsymbol{\mu}_x\right)^T \Sigma^{-1} \left(\mathbf{x}_k - \boldsymbol{\mu}_x\right) \tag{331}$$

$$t = \frac{z_{ek}}{\sqrt{1-h_{kk}}} \tag{332}$$

4.5.4. 新たなデータの評価

$$D^2 = \left(\mathbf{x} - \boldsymbol{\mu}_x\right)^T \Sigma^{-1} \left(\mathbf{x} - \boldsymbol{\mu}_x\right) \tag{333}$$

$$\hat{a}_0 + \boldsymbol{\alpha}^T \mathbf{x} \hat{a}_1 - t_p\left(\phi_e; P\right) \sqrt{\left(\frac{1}{n} + \frac{D^2}{n}\right) s_e^{(2)}} \le Y \le \hat{a}_0 + \boldsymbol{\alpha}^T \mathbf{x} \hat{a}_1 + t_p\left(\phi_e; P\right) \sqrt{\left(\frac{1}{n} + \frac{D^2}{n}\right) s_e^{(2)}} \tag{334}$$

$$\hat{a}_0 + \boldsymbol{\alpha}^T \mathbf{x} - t\left(\phi_e, P\right) \sqrt{\left(1 + \frac{1}{n} + \frac{D^2}{n}\right) s_e^{(2)}} \le y \le \hat{a}_0 + \boldsymbol{\alpha}^T \mathbf{x} + t\left(\phi_e, P\right) \sqrt{\left(1 + \frac{1}{n} + \frac{D^2}{n}\right) s_e^{(2)}} \tag{335}$$

4.6. まとめ

この章の結果をまとめる。

m 個からなる説明変数の回帰直線は以下になると仮定する。

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_m X_m$$

この係数は以下のように決定される。

$$\begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \cdots \\ \hat{a}_m \end{pmatrix} = \begin{pmatrix} S_{11}^{(2)} & S_{12}^{(2)} & \cdots & S_{1m}^{(2)} \\ S_{21}^{(2)} & S_{22}^{(2)} & \cdots & S_{2m}^{(2)} \\ \cdots & \cdots & \ddots & \cdots \\ S_{m1}^{(2)} & S_{m2}^{(2)} & \cdots & S_{mm}^{(2)} \end{pmatrix}^{-1} \begin{pmatrix} S_{1y}^{(2)} \\ S_{2y}^{(2)} \\ \cdots \\ S_{py}^{(2)} \end{pmatrix}$$

回帰の精度は以下の係数で評価される。

$$R^2 = \frac{S_r^{(2)}}{S_{yy}^{(2)}}$$

ここで$S_{yy}^{(2)}$ と $S_r^{(2)}$は以下で定義される分散である。

$$S_{yy}^{(2)} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$$

$$S_r^{(2)} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{y}\right)^2$$

回帰の有効性は以下のF値で評価される。

$$F = \frac{s_r^{(2)}}{s_e^{(2)}}$$

ここで、F 値の中の不偏分散は以下のように定義される。

$$s_r^{(2)} = \frac{nS_r^{(2)}}{m}$$

$$s_e^{(2)} = \frac{nS_e^{(2)}}{n-\left(m+1\right)}$$

分散 $S_e^{(2)}$ は以下で与えられる。

$$S_e^{(2)} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - Y_i\right)^2$$

我々は推定確率P におけるF 値、すなわち$F_p\left(m, n-\left(m+1\right)\right)$も評価し、以下のように回帰の有効性を判定する。

$$\begin{cases} F \le F_p\left(m, n-\left(m+1\right)\right) \Rightarrow invalid \\ F > F_p\left(m, n-\left(m+1\right)\right) \Rightarrow valid \end{cases}$$

各データの残差誤差は以下のように評価される。

$$t_k = \frac{z_{ek}}{\sqrt{1-h_{kk}}}$$

ここで各パラメータは以下である。

$$z_{ek} = \frac{e_k}{\sqrt{s_e^{(2)}}}$$

$$h_{kk} = \frac{1}{n} + \frac{D_k^2}{n}$$

さらに、

$$D_k^2 = \begin{pmatrix} x_{k1} - \overline{x}_1 & x_{k2} - \overline{x}_2 & \cdots & x_{km} - \overline{x}_m \end{pmatrix} \begin{pmatrix} S^{11(2)} & S^{12(2)} & \cdots & S^{1m(2)} \\ S^{21(2)} & S^{22(2)} & \cdots & S^{2m(2)} \\ \cdots & \cdots & \ddots & \cdots \\ S^{m1(2)} & S^{m2(2)} & \cdots & S^{mm(2)} \end{pmatrix} \begin{pmatrix} x_{k1} - \overline{x}_1 \\ x_{k2} - \overline{x}_2 \\ \cdots \\ x_{km} - \overline{x}_m \end{pmatrix}$$

この D_k^2 はマハラノビスの距離の 2 乗と呼ばれる。.

回帰の誤差範囲は以下で与えられる。

$$\hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2 + \cdots + \hat{a}_m x_m - t_p(\phi_e; P)\sqrt{\left(\frac{1}{n} + \frac{D^2}{n}\right) s_e^{(2)}}$$

$$\leq Y \leq$$

$$\hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2 + \cdots + \hat{a}_m x_m + t_p(\phi_e; P)\sqrt{\left(\frac{1}{n} + \frac{D^2}{n}\right) s_e^{(2)}}$$

目的変数の推定範囲は以下で与えられる。

$$\hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2 + \cdots + \hat{a}_m x_m - t(\phi_e, P)\sqrt{\left(1 + \frac{1}{n} + \frac{D^2}{n}\right) s_e^{(2)}}$$

$$\leq y \leq$$

$$\hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2 + \cdots + \hat{a}_m x_m + t(\phi_e, P)\sqrt{\left(1 + \frac{1}{n} + \frac{D^2}{n}\right) s_e^{(2)}}$$

変数 1 の純粋な目的変数に対する依存性は部分相関係数 $r_{(1y),(2,3,\cdots,m)}$ によって評価される。

それは、以下で与えられる。

$$r_{(1y),(2,3,\cdots,m)} = \frac{S_{uv_1}^{(2)}}{\sqrt{S_{v_1v_1}^{(2)}}\sqrt{S_{uu}^{(2)}}}$$

ここで、u,v は他の変数の影響を取り除いた説明変数、目的変数に相当する。それらは、以下で与えられる

$$u_k = y_k - Y_k$$

$$v_{k1} = x_{k1} - X_{k1}$$

ここで、Y_k 、X_k は他の変数で表現される回帰直線であり、以下で与えられる。

$$Y_k = c_0 + c_2 x_{k2} + c_3 x_{k3} + \cdots + c_m x_{km}$$

$$X_{k1} = d_0 + d_2 x_{k2} + d_3 x_{k3} + \cdots + d_m x_{km}$$

係数 $c_2, c_3, \cdots, c_m$ は以下から決定される。

$$\begin{pmatrix} c_2 \\ c_3 \\ \cdots \\ c_m \end{pmatrix} = \begin{pmatrix} S_{22}^{(2)} & S_{23}^{(2)} & \cdots & S_{2m}^{(2)} \\ S_{32}^{(2)} & S_{33}^{(2)} & \cdots & S_{3m}^{(2)} \\ \cdots & \cdots & \ddots & \cdots \\ S_{m2}^{(2)} & S_{m3}^{(2)} & \cdots & S_{mm}^{(2)} \end{pmatrix}^{-1} \begin{pmatrix} S_{2y}^{(2)} \\ S_{3y}^{(2)} \\ \cdots \\ S_{my}^{(2)} \end{pmatrix}$$

c_0 は以下から決定される。

$$\bar{y} = c_0 + c_2 \bar{x}_2 + c_3 \bar{x}_3 + \cdots + c_m \bar{x}_m$$

係数 $d_2, d_3, \cdots, d_m$ は以下から決定される

$$\begin{pmatrix} d_2 \\ d_3 \\ \cdots \\ d_m \end{pmatrix} = \begin{pmatrix} S_{22}^{(2)} & S_{23}^{(2)} & \cdots & S_{2m}^{(2)} \\ S_{32}^{(2)} & S_{33}^{(2)} & \cdots & S_{3m}^{(2)} \\ \cdots & \cdots & \ddots & \cdots \\ S_{m2}^{(2)} & S_{m3}^{(2)} & \cdots & S_{mm}^{(2)} \end{pmatrix}^{-1} \begin{pmatrix} S_{21}^{(2)} \\ S_{31}^{(2)} \\ \cdots \\ S_{m1}^{(2)} \end{pmatrix}$$

d_0 は以下から決定される。

$$\bar{x}_1 = d_0 + d_2 \bar{x}_2 + d_3 \bar{x}_3 + \cdots + d_m \bar{x}_m$$

有効な説明変数の選択を以下のように行う。

説明変数なしの状態からスタートする。それをモデル 0 とする。 その回帰は以下となる。

$$Model0 : Y_i = \bar{y}$$

対応する誤差分散 $S_{e(M0)}^{(2)}$ は以下で与えられる。

$$S_{e(M0)}^{(2)} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - Y_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = S_{yy}^{(2)}$$

次のステップでは、一つの説明変数を選択する。それをモデル 1 とする。

その変数を使った回帰直線は x 座標 x_{1l} を使って以下で与えられる。

$$Y_i = a_0 + a_1 x_{il_1}$$

対応する誤差分散 $S_{e(M1)}^{(2)}$ は以下で与えられる。

$$S_{e(M1)}^{(2)} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - Y_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \left(a_0 + a_1 x_{il_1}\right)\right]^2$$

ここで、以下のパラメータを評価する。

$$F_1 = \frac{\left(nS_{e(M0)}^{(2)} - nS_{e(M1)}^{(2)}\right)\Big/\left(\phi_{e(M0)} - \phi_{e(M1)}\right)}{nS_{e(M1)}^{(2)}\Big/\phi_{e(M1)}}$$

これは、$F\left(\phi_{e(M0)} - \phi_{e(M1)}, \phi_{e(M1)}\right)$ 分布に従う。ここで、$\phi_{e(M0)}$ と $\phi_{e(M1)}$ はそれぞれの自由度であり、以下で与えられる。

$$\phi_{e(M0)} = n - 1$$

$$\phi_{e(M1)} = n - 2$$

この説明変数の有効性は以下で評価する。

$$\begin{cases} F_1 \geq F\left(\phi_{e(M0)} - \phi_{e(M1)}, \phi_{e(M1)}\right) & valid \\ F_1 < F\left(\phi_{e(M0)} - \phi_{e(M1)}, \phi_{e(M1)}\right) & invalid \end{cases}$$

上の条件を満足するものがあれば、この中で、最高の値をとる変数を選択する。
このプロセスを以下のように繰り返す。

説明変数を選択し、以下のパラメータを評価する。

$$F_{i+1} = \frac{\left(nS_{e(Mi)}^{(2)} - nS_{e(Mi+1)}^{(2)}\right)\Big/\left(\phi_{e(Mi)} - \phi_{e(Mi+1)}\right)}{nS_{e(Mi+1)}^{(2)}\Big/\phi_{e(MI+1)}}$$

これは、$F\left(\phi_{e(Mi)} - \phi_{e(Mi+1)}, \phi_{e(Mi+1)}\right)$ 分布にしたがう。ここで、$\phi_{e(Mi)}$ と $\phi_{e(Mi+1)}$ はそれぞれの自由度で、以下で与えられる。

$$\phi_{e(Mi)} = n - i - 1$$

$$\phi_{e(Mi+1)} = n - i - 2$$

もし、上の条件を満足する説明変数があれば、その中で最高の値をとつ変数を選択する。もしも、条件を満足する変数がなければ、このプロセスを終了する。

もう一つの変数の選択方法を示す。

最初に一つの変数を選択し、それぞれの変数の場合の R を評価する。その中で、最大のものを与える変数を選択する。その変数を l_1 とする。

次に変数 l_1 と残りの変数の中の一つの変数を選択し、その R を評価する。そして、最大値を与える変数をを選択する。その後、全ての係数について、以下を評価する。

$$F = \frac{\hat{a}_i}{\frac{S^{ii(2)}}{n} s_e^{(2)}}$$

これと $F_c(1, n-2-1; P)$ を比べる。(通常は P として 0.86 が採用される。)もし、すべての係数について $F \geq F_c$ であれば、その変数を採用し、そうでなければ不採用にする。このプロセスを繰り返す。

重回帰の係数は以下の行列で表現される。

$$\boldsymbol{\beta} = \left(X^T X\right)^{-1} X^T Y$$

ここで、各点は以下で表現される。

以下の行列とベクトルを用いる。

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$X = \begin{pmatrix} 1 & x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1p} - \bar{x}_p \\ 1 & x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{np} - \bar{x}_p \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix}$$

これを用いて、データ点は以下のように表現される。

$$y_i = a_0 + a_1\left(x_{i1} - \bar{x}_1\right) + a_2\left(x_{i2} - \bar{x}_2\right) + \cdots + a_p\left(x_{ip} - \bar{x}_p\right) + e_i$$

ただし、以下である。

$$\mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$