# 多変量ゼミ　§1.12 回帰式による予測値の区間推定（重回帰の場合）

髙見澤 真央

ここでは，単回帰の流れをもとに，重回帰における目的変数の予測値の区間推定を行う．

おおまかな流れとして，

(1)重回帰モデルによる目的変数の期待値および分散を計算する

(2) (1)で得られた結果から，信頼区間をわりだし，区間推定を行う

このように進めていく．

また，前節の§1.11 回帰係数の推定と検定(重回帰の場合)の範囲で求めた式を参照する場合は，式番号に☆のマークを付けて表す．

## 1. 回帰式による予測値の区間推定(重回帰の場合)

### 1.1. 予測値の期待値と分散

説明変数$(x_1,\cdots,x_p)$がある特定の値$(x_{10},\cdots,x_{p0})$をとるときの目的変数$y$の期待値$\eta_0$は，回帰式を用いて，

$$Y_0=\hat{a}_0+\hat{a}_1x_{10}+\cdots+\hat{a}_px_{p0} \tag{1}$$

によって推定(予測)される．このとき，$Y_0$の期待値は，

$$\begin{aligned}\mathrm{E}(Y_0)&=\mathrm{E}\left(\hat{a}_0+\hat{a}_1x_{10}+\cdots+\hat{a}_px_{p0}\right)\\&=\mathrm{E}(\hat{a}_0)+\mathrm{E}(\hat{a}_1)x_{10}+\cdots+\mathrm{E}\left(\hat{a}_p\right)x_{p0}\\&=a_0+a_1x_{10}+\cdots+a_px_{p0}\\&\qquad\text{（☆式(8),(11)}\ \mathrm{E}\left(\hat{a}_j\right)=a_j,\mathrm{E}(\hat{a}_0)=a_0\ \text{より）}\\&=\eta_0\end{aligned} \tag{2}$$

となり，予測値$Y_0$は$\eta_0$ に対する不定推定値であることがわかる．

また，$Y_0$の分散は，

$$\begin{aligned}\mathrm{V}(Y_0)&=\mathrm{E}[\{Y_0-\mathrm{E}(Y_0)\}^2]\\&=\mathrm{E}\left[\left\{\left(\hat{a}_0+\hat{a}_1x_{10}+\cdots+\hat{a}_px_{p0}\right)-\left(a_0+a_1x_{10}+\cdots+a_px_{p0}\right)\right\}^2\right]\\&\qquad\text{（式(1)}\ Y_0=\hat{a}_0+\hat{a}_1x_{10}+\cdots+\hat{a}_px_{p0}\ \text{より）}\\&=\mathrm{E}\left[\left\{\left(\hat{a}_0+\sum_{j=1}^{p}\hat{a}_jx_{j0}\right)-\left(a_0+\sum_{j=1}^{p}a_jx_{j0}\right)\right\}^2\right]\\&=\mathrm{E}\left[\left\{(\hat{a}_0-a_0)+\sum_{j=1}^{p}x_{j0}\left(\hat{a}_j-a_j\right)\right\}^2\right]\\&=\mathrm{E}\left[(\hat{a}_0-a_0)^2+2\sum_{j=1}^{p}x_{j0}(\hat{a}_0-a_0)\left(\hat{a}_j-a_j\right)+\sum_{j=1}^{p}\sum_{l=1}^{p}x_{j0}x_{l0}\left(\hat{a}_j-a_j\right)(\hat{a}_l-a_l)\right]\\&=\mathrm{E}[(\hat{a}_0-a_0)^2]+2\sum_{j=1}^{p}x_{j0}\,\mathrm{E}\left[(\hat{a}_0-a_0)\left(\hat{a}_j-a_j\right)\right]+\sum_{j=1}^{p}\sum_{l=1}^{p}x_{j0}x_{l0}\,\mathrm{E}\left[\left(\hat{a}_j-a_j\right)(\hat{a}_l-a_l)\right]\end{aligned}$$

$$
\begin{aligned}
\mathrm{V}(Y_0) &= \mathrm{V}(\hat{a}_0) + 2\sum_{j=1}^{p} x_{j0}\mathrm{Cov}(\hat{a}_0, \hat{a}_j) + \sum_{j=1}^{p}\sum_{l=1}^{p} x_{j0}x_{l0}\,\mathrm{Cov}(\hat{a}_j, \hat{a}_l) \\
&= \frac{\left(1 + \sum_{\mathrm{j=1}}^{\mathrm{p}}\sum_{\mathrm{l=1}}^{\mathrm{p}} \bar{x}_j\bar{x}_l\, s^{jl}\right)\sigma^2}{n} + 2\sum_{j=1}^{p} x_{j0}\left(-\frac{\sum_{l=1}^{p} \bar{x}_l\, s^{jl}\sigma^2}{n}\right) + \sum_{j=1}^{p}\sum_{l=1}^{p} x_{j0}x_{l0}\frac{s^{jl}\sigma^2}{n}
\end{aligned}
$$

（☆式(10) $\mathrm{Cov}(\hat{a}_j, \hat{a}_l) = \frac{s^{jl}\sigma^2}{n}$，　☆式(12)$\mathrm{V}(\hat{a}_0) = \left(\frac{1}{n} + \sum_{j=1}^{p}\sum_{l=1}^{p}\frac{\bar{x}_j\bar{x}_l s^{jl}}{n}\right)\sigma^2$

☆式(13) $\mathrm{Cov}(\hat{a}_0, \hat{a}_j) = -\sum_{l=1}^{p}\frac{\bar{x}_l s^{jl}\sigma^2}{n}$ ）

$$
\begin{aligned}
&= \left\{\frac{1}{n} + \frac{1}{n}\left(\sum_{j=1}^{p}\sum_{l=1}^{p} \bar{x}_j\bar{x}_l\, s^{jl} - 2\sum_{j=1}^{p}\sum_{l=1}^{p} x_{j0}\bar{x}_l s^{jl} + \sum_{j=1}^{p}\sum_{l=1}^{p} x_{j0}x_{l0}\, s^{jl}\right)\right\}\sigma^2 \\
&= \left\{\frac{1}{n} + \frac{1}{n}\sum_{j=1}^{p}\sum_{l=1}^{p}(\bar{x}_j\bar{x}_l - x_{j0}\bar{x}_l - x_{l0}\bar{x}_j + x_{j0}x_{l0})s^{jl}\right\}\sigma^2
\end{aligned}
$$

( $2\sum_{j=1}^{p}\sum_{l=1}^{p} x_{j0}\bar{x}_l = x_{j0}\bar{x}_l + x_{l0}\bar{x}_j$ )

$$
= \left\{\frac{1}{n} + \frac{1}{n}\sum_{j=1}^{p}\sum_{l=1}^{p}(x_{j0} - \bar{x}_j)(x_{l0} - \bar{x}_l)s^{jl}\right\}\sigma^2 \tag{3}
$$

前節において，回帰係数$\hat{a}_j$と定数項$\hat{a}_0$は，正規分布に従うがわかった．

$$
\hat{a}_j = \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{p} s^{jl}(x_{li} - \bar{x}_l)y_i
$$

$$
\hat{a}_0 = \frac{1}{n}\sum_{i=1}^{n}\left\{1 - \sum_{j=1}^{p}\sum_{l=1}^{p} s^{jl}(x_{li} - \bar{x}_l)\bar{x}_j\right\}y_i
$$

いずれも正規分布に従う変数$y_i$の 1 次式で表される ⇒ 回帰係数$\hat{a}_j$と定数項$\hat{a}_0$も正規分布に従う

$\hat{a}_0, \hat{a}_1, \cdots, \hat{a}_p$の 1 次式である式(1) $Y_0 = \hat{a}_0 + \hat{a}_1 x_{10} + \cdots + \hat{a}_p x_{p0}$ もまた，正規分布に従うことがわかる．ここで，$Y_0$を標準化すると，

$$
\begin{aligned}
u &= \frac{Y_0 - \eta_0}{\sqrt{\mathrm{V}(Y_0)}} \\
&= \frac{Y_0 - \eta_0}{\sqrt{\left\{\frac{1}{n} + \frac{1}{n}\sum_{\mathrm{j=1}}^{\mathrm{p}}\sum_{\mathrm{l=1}}^{\mathrm{p}}(x_{j0} - \bar{x}_j)(x_{l0} - \bar{x}_l)s^{jl}\right\}\sigma^2}}
\end{aligned}
$$

となる．$u$は標準正規分布$N(0,1)$に従う．ただし，このとき未知の誤差分散$\sigma^2$を含んでおり，これを不偏推定値$V_e = \frac{F(\hat{a}_0, \hat{a}_1, \cdots, \hat{a}_p)}{n-p-1}$ で置き換えて得られる統計量は，

$$
t = \frac{Y_0 - \eta_0}{\sqrt{\left\{\frac{1}{n} + \frac{1}{n}\sum_{\mathrm{j=1}}^{\mathrm{p}}\sum_{\mathrm{l=1}}^{\mathrm{p}}(x_{j0} - \bar{x}_j)(x_{l0} - \bar{x}_l)s^{jl}\right\}V_e}} \tag{4}
$$

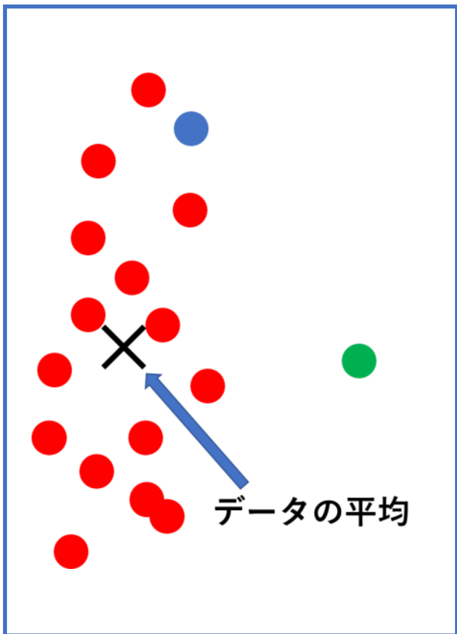
ここで，

$$D_0^2 = \sum_{j=1}^{P}\sum_{l=1}^{P}(x_{j0} - \bar{x}_j)(x_{l0} - \bar{x}_l)s^{jl} \tag{5}$$

と置くとする．この$D_0^2$は，単回帰における

$$t = \frac{Y_0 - \eta_0}{\sqrt{\left\{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{ns_{xx}}\right\}V_e}}$$

この式の分母の$(x_0 - \bar{x})^2/s_{xx}$(=標準偏差で標準化した平均からの距離の 2 乗)を，多変量の場合に一般化したものに相当し，点$(x_{10}, \cdots, x_{p0})$と重心$(\bar{x}_1, \cdots, \bar{x}_p)$との間のマハラノビスの汎距離と呼ばれる．




ざっくり マハラノビスの距離とは・・・

左の図において，赤い点のデータ群があるときに，
青い点と緑の点はどちらの点がよりデータ群から離れているかを判断したい．

データの平均からのユークリッド距離は，ほぼ同じであるが，
点の散らばりを見るとと，緑の点の方があきらかに赤い点のデータ群から離れていることがわかる．
正しく判断するには，データの散らばり具合も考慮した「データ群からの距離」を考える必要があり，これを実現しているのがマハラノビスの距離である．

主に判別分析や異常検知の分野で使われている．

式(5) $D_0^2$を用いると，式(4)は次のようになる．

$$t = \frac{Y_0 - \eta_0}{\sqrt{\left\{\frac{1}{n} + \frac{1}{n}D_0^2\right\}V_e}} \tag{6}$$

このとき，$t$ は自由度$n - p - 1$の$t$分布に従う．

信頼区間は，平均して 100 回中 100$(1 - \alpha)$回 母集団の$a_0, a_j$ を含むことが保証された範囲のこと．

これから，$\eta_0$に対する信頼率$1 - \alpha$の信頼区間は，

$$-t_\alpha(n - p - 1) \le \frac{Y_0 - \eta_0}{\sqrt{\left(\frac{1}{n} + \frac{1}{n}D_0^2\right)V_e}} \le t_\alpha(n - p - 1)$$

すなわち，

$$Y_0 - t_\alpha(n - p - 1)\sqrt{\left(\frac{1}{n} + \frac{1}{n}D_0^2\right)V_e} \le \eta_0 \le Y_0 + t_\alpha(n - p - 1)\sqrt{\left(\frac{1}{n} + \frac{1}{n}D_0^2\right)V_e} \tag{7}$$

このように求められる．式(7)をみると，$D_0^2$の小さいところ，すなわち重心の近くでは信頼区間の幅が小さく，逆に$D_0^2$ が大きく，重心から遠く離れたところでは，信頼区間の幅が大きくなることがわかる．

## 1.2. 目的変数 y の信頼区間と区間推定

次に，説明変数の特定の値$(x_{10},\cdots,x_{p0})$に対して観測される目的変数$y$の値の信頼区間について考える．$y$は，これまでに観測されている$(y_1,\cdots,y_n)$やこれらにもとづく$\hat{a}_1,\cdots,\hat{a}_p$とは独立に，平均$\eta_0$，分散$\sigma^2$の正規分布に従うため，$Y_0-y$の期待値と分散は，

$$\begin{aligned}\mathrm{E}(Y_0-y)&=\mathrm{E}(Y_0)-\mathrm{E}(y)\\&=\eta_0-\eta_0\quad\text{（式(2)}\ \mathrm{E}(Y_0)=\eta_0\ \text{より）}\\&=0\end{aligned}\tag{8}$$

$$\begin{aligned}\mathrm{V}(Y_0-y)&=\mathrm{E}[(Y_0-y)^2]\\&=\mathrm{E}[\{(Y_0-\eta_0)-(y-\eta_0)\}^2]\\&=\mathrm{E}[(Y_0-\eta_0)^2]+\mathrm{E}[(y-\eta_0)^2]-2\mathrm{E}[(Y_0-\eta_0)(y-\eta_0)]\\&=\mathrm{V}(Y_0)+\mathrm{V}(y)-2\mathrm{Cov}(Y_0,y)\\&=\mathrm{V}(Y_0)+\mathrm{V}(y)\quad\text{（}Y_0\text{と}y\text{は無相関のため，}2\mathrm{Cov}(Y_0,y)=0\text{）}\\&=\left(\frac{1}{n}+\frac{1}{n}D_0^2\right)\sigma^2+\sigma^2\\&\text{（式(3)}\ \mathrm{V}(Y_0)=\left\{\frac{1}{n}+\frac{1}{n}\sum_{j=1}^{p}\sum_{l=1}^{p}(x_{j0}-\bar{x}_j)(x_{l0}-\bar{x}_l)s^{jl}\right\}\sigma^2\ \text{より）}\\&=\left(1+\frac{1}{n}+\frac{1}{n}D_0^2\right)\sigma^2\end{aligned}\tag{9}$$

となる．これを用いて，$Y_0-y$を標準化し，$\sigma^2$を不偏推定値$V_e=\frac{F(\hat{a}_0,\hat{a}_1,\cdots,\hat{a}_p)}{n-p-1}$で置き換えると，

$$t'=\frac{Y_0-y}{\sqrt{\left(1+\frac{1}{n}+\frac{1}{n}D_0^2\right)V_e}}\tag{10}$$

を得る．$t'$は自由度$n-p-1$の$t$分布に従う．

これより，$(x_{10},\cdots,x_{p0})$に対応して観測される目的変数$y$に対する信頼率$1-\alpha$の信頼区間は，

$$-t_\alpha(n-p-1)\le\frac{Y_0-y}{\sqrt{\left(1+\frac{1}{n}+\frac{1}{n}D_0^2\right)V_e}}\le t_\alpha(n-p-1)$$

すなわち，

$$Y_0-t_\alpha(n-p-1)\sqrt{\left(1+\frac{1}{n}+\frac{1}{n}D_0^2\right)V_e}\le y\le Y_0+t_\alpha(n-p-1)\sqrt{\left(1+\frac{1}{n}+\frac{1}{n}D_0^2\right)V_e}\tag{11}$$

このように与えられる．

## 2. 説明変数の選択

これまでの議論では，回帰モデルに含まれる説明変数$x_1, \cdots, x_p$は定められたものとして計算してきた．しかし，実際の現象を分析する場合には，目的変数$y$に影響を及ぼす可能性のある数多の変数の中から，次のようなことを考慮し，変数を取りあげるのが一般的である．

1. 回帰モデルに無駄な変数(真の回帰係数が 0 であるような変数)が含まれる場合，
   誤差分散の推定値$V_e$の自由度$n-p-1$が小さくなり，回帰係数$\hat{a}_j$や予測値$Y_0$の推定制度が悪くなる．

2. 必要な変数(真の回帰係数が 0 でない変数)が回帰モデルから外れている場合，
   回帰係数の推定値や目的変数の予測値が偏りを持ち，誤差分散の推定値$V_e$は過大評価になる．

3. 説明変数の中に互いに相関が高い変数が含まれる場合，
   分散共分散行列$V=\left(s_{jl}\right)$の行列式が 0 に近くなるため，逆行列の要素$s^{jj}=V_{jj}/|V|$が大きくなり，
   回帰係数の推定精度が悪くなる．特にある説明変数と残りの変数の重相関係数が $R=1$ のとき，
   $V$の行列式は 0 となり，逆行列が存在しない．そのため回帰係数$\hat{a}_j$を求めることができない．
   これを多重共線性の問題という．

多重共線性とは・・・
　重回帰分析を行ったとき，互いに関連性の高い説明変数が存在すると，解析上の計算が不安定となり，回帰式の精度が極端に悪くなる現象．またの名をマルチコ現象と呼ぶ．

### 2.1. 総あたり法

$p$個の説明変数の候補の中から，1~$p$個の変数で考えられるすべての組合せ $2^p-1$ 通りの回帰モデルを検討する方法．最も単純な方法ではあるが，候補となる変数の個数$p$が多くなると，組合せの数が急速に大きくなり，計算時間が膨大になる．

### 2.2. 前進選択法

説明変数が 1 つも含まれない状態からスタートして，次のような手順で変数を 1 つずつ増加させる．

(ⅰ) 目的変数$y$との単相関が最大(言いかえると，1 つずつ順に変数を採用してみて回帰式を計算したとき，回帰係数検定のための $t$ の絶対値または$F$値が最大)の変数を選び，回帰係数が 0 であるという仮説の検定を行い，仮説が棄却されなければ，どの変数もモデルに含めない．仮説が棄却されれば，この変数をモデルにとりこみ次のステップ(ⅱ)へ進む．

(ⅱ) 既に入っている変数に加えて残りの変数を 1 つずつ順に採用してみて，偏相関係数が最大(回帰係数検定のための $t$ の絶対値または$F$値が最大)の変数を選ぶ．選ばれた変数に対する回帰係数が 0 であるという仮説の検定を行い，仮説が棄却されなければ終了．仮説が棄却されれば，変数を取り込んで次のステップ(ⅲ)へ進む．

(ⅲ) 回帰式を計算する．もしモデルにすべての変数が含まれていれば終了．
そうでなければステップ(ⅱ)へ戻る．

## 2.3. 後退消去法

説明変数の候補がすべて含まれた状態からスタートして，次のような手順で変数を 1 つずつ減少させる．

（ｉ）モデルに含まれる各変数に対する回帰係数検定のための $t$ または $F$ 値を計算し，絶対値が最小となる変数を選ぶ．回帰係数が 0 であるという仮説が棄却されなければ，その変数を落として次のステップ(ⅱ)へ進む．棄却されれば終了．

（ⅱ）もしモデルに含まれる変数がなくなっていれば終了．そうでなければ回帰式を計算しなおして，ステップ(ｉ)へ戻る．

## 2.4. 逐次法

前進選択法では，1 度モデルに含まれた変数が途中で落とされることはなかった．この点を改良し，次のような手順で変数を増減させる．

（ｉ）目的変数との単相関が最大の変数を選ぶ．選ばれた変数に対する回帰係数が 0 であるという仮説の検定を行い，棄却されなければどの変数も回帰モデルに含めない．棄却されれば，この変数をとりこんで次のステップ(ⅱ)へ進む．

（ⅱ）既に入っている変数に加えて残りの変数を 1 つずつ順に採用してみて，偏相関係数が最大の変数を選ぶ．選ばれた変数に対する回帰係数が 0 であるという仮説の検定を行い，仮説が棄却されなければ終了．仮説が棄却されれば，変数を取り込んで次のステップ(ⅲ)へ進む．

（ⅲ）回帰式を計算し，各変数について回帰係数検定を行い，$F$値が最小になる変数について，仮説が棄却されなければ，その変数を落とす．

（ⅳ）すべての変数がとりこまれていれば終了．そうでなければステップ(ⅱ)に戻る．

上の 4 つの各方法において，回帰係数の検定は次のように行う．

$p$個の変数を含むモデルでの変数$x_j$に対する回帰係数が 0 という仮説$\mathrm{H_0}: a_j = 0$の検定は，
前節の式(20)

$$|t| = \frac{\left|\hat{a}_j - a_j^{(0)}\right|}{\sqrt{\frac{s^{jj}V_e}{n}}} \geq t_\alpha(n-p-1)$$

で，$a_j^{(0)} = 0$とおき，

$$t = \frac{\hat{a}_j}{\sqrt{s^{jj}V_e/n}} \tag{12}$$

の値を求めて，自由度$n-p-1$の$t$分布の限界値と比較し，$|t| \geq t_\alpha(n-p-1)$ならば仮説を棄却，
$|t| < t_\alpha(n-p-1)$ならば仮説を採択する．

また，$t$分布と$F$分布の関係，
「ある$x$が自由度$n$の$t$分布に従うとき，$x^2$は自由度$(1,n)$の$F$分布に従う」ことから，

$$F = \frac{\hat{a}_j^2}{s^{jj}V_e/n} \tag{13}$$

の値を求めて，自由度$(1, n-p-1)$の$F$分布の限界値と比較し，$|F| \geq F_{n-p-1}^{1}(\alpha)$ならば仮説を棄却する，ことで同じ結果が得られる．

補足 多重共線性

説明変数が 2 つの回帰モデル $y = \hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2$ を例に，多重共線性の問題について考える．回帰係数$\hat{a}_1$を求める式は，§1.5 の線形重回帰の範囲から，

$$\hat{a}_1 = \frac{\begin{vmatrix} s_{y1} & s_{12} \\ s_{y1} & s_{22} \end{vmatrix}}{\begin{vmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{vmatrix}} \tag{14}$$

である．この式の分母，すなわち，分散共分散行列$V = (s_{jl})$の行列式は，

$$\begin{aligned} \text{（式(14)の分母）} \quad &= s_{11}s_{22} - s_{12}s_{21} \\ &= s_{11}s_{22}\left(1 - \frac{s_{12}}{\sqrt{s_{11}s_{22}}}\frac{s_{21}}{\sqrt{s_{11}s_{22}}}\right) \\ &= s_{11}s_{22}(1 - r^2) \end{aligned} \tag{15}$$

となる．このとき $r$ は，$x_1$と$x_2$の相関係数である．

$x_1$と$x_2$の相関が高い，すなわち相関係数 $r$ が$\pm 1$に近い値をとるとき，分散共分散行列の行列式は，0 に近くなり，回帰係数$\hat{a}_1$は式(14)の分子の行列式$\begin{vmatrix} s_{y1} & s_{12} \\ s_{y1} & s_{22} \end{vmatrix}$の変化に大きく影響されることがわかる．変数値の変化によって，回帰係数の推定値が大きく変わるため，係数推定値の分散が大きくなり，推定結果の信頼性がおちる．また，特に相関係数が$r = \pm 1$（線形従属)のとき，分散共分散行列の行列式は 0 となり，それを分母にもつ式(14)は計算ができず，そもそも回帰係数を求めることができない．

重回帰分析では，候補となる説明変数の間に相関がないことを確認し，相関が見られた場合にはその説明変数を外すことで，多重共線性の問題を回避する必要がある．