§1.12 回帰式による予測値の区間推定 (重回帰の場合)
§1.13 説明変数の選択

富島諒

2021 年 6 月 17 日

§1.11 で紹介した数式は式番号の前に ♣ を付ける.

1 §1.12 回帰式による予測値の区間推定 (重回帰の場合)

本節の目的は,

1. 重回帰モデルによる目的変数の基本統計量を求める
2. 基本統計量を元に信頼区間を算出し, 区間推定を行う

である.

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 η_0 は, 回帰式を用いて

$$Y_0 = \hat{a}_0 + \hat{a}_1 x_{10} + \cdots + \hat{a}_p x_{p0} \tag{1}$$

によって, 推定 (予測) することができる. このとき, Y_0 の期待値は

$$\begin{aligned} \mathrm{E}(Y_0) &= \mathrm{E}(\hat{a}_0 + \hat{a}_1 x_{10} + \cdots + \hat{a}_p x_{p0}) \\ &= \mathrm{E}(\hat{a}_0) + \mathrm{E}(\hat{a}_1) x_{10} + \cdots + \mathrm{E}(\hat{a}_p) x_{p0} \\ &= a_0 + a_1 x_{10} + \cdots + a_p x_{p0} = \eta_0 \\ &\quad (\because \clubsuit \text{ 式 (11a) } \mathrm{E}(\hat{a}_j) = a_j, \text{ 式 (11d) } \mathrm{E}(\hat{a}_0) = a_0 \text{ より}) \end{aligned} \tag{2}$$

となり, 予測値 Y_0 は η_0 に対する不偏推定値である. また Y_0 の分散は式 (3) のようになる.

$$
\begin{aligned}
\mathrm{V}(Y_0) &= \mathrm{E}\left[\{Y_0 - \mathrm{E}(Y_0)\}^2\right] \\
&= \mathrm{E}\left[\{(\hat{a}_0 + \hat{a}_1 x_{10} + \cdots + \hat{a}_p x_{p0}) - (a_0 + a_1 x_{10} + \cdots + a_p x_{p0})\}^2\right] \\
&\quad (\because \text{式 (1)}\ \ Y_0 = \hat{a}_0 + \hat{a}_1 x_{10} + \cdots + \hat{a}_p x_{p0} \text{ より}) \\
&= \mathrm{E}\left[\left\{\left(\hat{a}_0 + \sum_{j=1}^{p} \hat{a}_j x_{j0}\right) - \left(a_0 + \sum_{j=1}^{p} a_j x_{j0}\right)\right\}^2\right] \\
&= \mathrm{E}\left[\left\{(\hat{a}_0 - a_0) + \sum_{j=1}^{p} x_{j0}(\hat{a}_j - a_j)\right\}^2\right] \\
&= \mathrm{E}\left[(\hat{a}_0 - a_0)^2 + 2\sum_{j=1}^{p} x_{j0}(\hat{a}_0 - a_0)(\hat{a}_j - a_j) + \sum_{j=1}^{p}\sum_{l=1}^{p} x_{j0}x_{l0}(\hat{a}_j - a_j)(\hat{a}_l - a_l)\right] \\
&= \mathrm{E}\left[(\hat{a}_0 - a_0)^2\right] + 2\sum_{j=1}^{p} x_{j0}\,\mathrm{E}\left[(\hat{a}_0 - a_0)(\hat{a}_j - a_j)\right] + \sum_{j=1}^{p}\sum_{l=1}^{p} x_{j0}x_{l0}\,\mathrm{E}\left[(\hat{a}_j - a_j)(\hat{a}_l - a_l)\right] \\
&= \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) \\
&= \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 + 2\sum_{j=1}^{p} x_{j0}\left(-\sum_{l=1}^{p} \frac{\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} \\
&\quad (\because \clubsuit \text{ 式 (11c)}\ \ \mathrm{Cov}(\hat{a}_j, \hat{a}_l) = \frac{s^{jl}\sigma^2}{n},\ \clubsuit \text{ 式 (11e)}\quad \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 \\
&\quad \clubsuit \text{ 式 (11f)}\quad \mathrm{Cov}(\hat{a}_0, \hat{a}_j) = -\sum_{l=1}^{p} \frac{\bar{x}_l s^{jl}\sigma^2}{n}) \\
&= \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 \\
&= \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 \qquad (3)
\end{aligned}
$$

ここで Y_0 を標準化すると,

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

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 σ^2 を含んでおり, これを不偏推

定値 $\mathrm{V}_e = \frac{\mathrm{F}_(\hat{a}_0,\hat{a}_1,\cdots,\hat{a}_p)}{n-p-1}$ で置き換えて得られる統計量は,

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

となる.

ここで,

$$\mathrm{D}_0^2 = \sum_{j=1}^{p}\sum_{l=1}^{p}(x_{j0} - \bar{x}_j)(x_{l0} - \bar{x}_l)s^{jl} \tag{5}$$

とする. この D_0^2 は単回帰における,

$$t = \frac{Y_0 - \eta_0}{\sqrt{\left\{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{ns_{xx}}\right\}\mathrm{V}_e}} \tag{6}$$

式 (6) 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) を重回帰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**[*1]と呼ばれる.

マハラノビスの汎距離

図の緑の点を赤色の点のグループか青色の点のグループに属するかを考える. このとき, 赤色のグループの平均点と青色のグループの平均点に関してユークリッド距離で比較すると, 赤色のグループの方が近いのでに緑色の点は赤色の区分される. しかし, 緑は青色に挟まれているので青色のグループに区分されるのが妥当だと考えられる. このように全てのデータの分布に応じた距離を定義するもののひとつがマハラノビス汎距離である.

図 1: マハラノビスの汎距離

式 (5) を用いると, 式 (4) は次のようになる.

$$t = \frac{Y_0 - \eta_0}{\sqrt{\left\{\frac{1}{n} + \frac{1}{n}\mathrm{D}_0^2\right\}\mathrm{V}_e}} \tag{7}$$

このとき, t は自由度 $n-p-1$ の t 分布に従い, η_0 に対する信頼率 $1-\alpha$ の信頼区間は,

$$-t_\alpha(n-p-1) \le \frac{Y_0 - \eta_0}{\sqrt{\left(\frac{1}{n} + \frac{1}{n}\mathrm{D}_0^2\right)\mathrm{V}_e}} \le t_\alpha(n-p-1)$$

すなわち,

$$Y_0 - t_\alpha(n-p-1)\sqrt{\left\{\frac{1}{n} + \frac{1}{n}\mathrm{D}_0^2\right\}\mathrm{V}_e} \le \eta_0 \le Y_0 + t_\alpha(n-p-1)\sqrt{\left\{\frac{1}{n} + \frac{1}{n}\mathrm{D}_0^2\right\}\mathrm{V}_e} \tag{8}$$

[*1] https://www.sist.ac.jp/~kanakubo/research/statistic/hanbetu_maha.html

のようになる. 式 (8) で, 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巾が小さく, D_0^2 の大きいところ, すなわち重心から遠く離れたところでは信頼区間の巾が大きくなることがわかる.

次に, 説明変数の特定の値 $(x_{10}, \cdots, x_{p0})$ に対応して得られるであろう目的変数 y の値の信頼区間について考える. y は今までに観測されている $(y_1, \cdots, y_n)$ やそれらに基づく $\hat{a}_1, \cdots, \hat{a}_p$ とは独立に平均 η_0, 分散 σ^2 の正規分布に従うため, $Y_0 - y$ の期待値と分散は,

$$\begin{aligned}
\mathrm{E}(Y_0 - y) &= \mathrm{E}(Y_0) - \mathrm{E}(y) \\
&= \eta_0 - \eta_0 \quad (\because \text{式 (2)}\ \ \mathrm{E}(Y_0) = \eta_0 \text{ より}) \\
&= 0
\end{aligned} \tag{9}$$

$$\begin{aligned}
\mathrm{V}(Y_0 - y) &= \mathrm{E}\left[(Y_0 - y)^2\right] \\
&= \mathrm{E}\left[\{(Y_0 - \eta_0) - (y - \eta_0)\}^2\right] \\
&= \mathrm{E}\left[(Y_0 - \eta_0)^2\right] + \mathrm{E}\left[(y - \eta_0)^2\right] - 2\,\mathrm{E}\left[(Y_0 - \eta_0)(y - \eta_0)\right] \\
&= \mathrm{V}(Y_0) + \mathrm{V}(y) - 2\,\mathrm{Cov}(Y_0, y) \\
&= \mathrm{V}(Y_0) + \mathrm{V}(y) \quad (\because\ Y_0 \text{ と } y \text{ は無相関}) \\
&= \left(\frac{1}{n} + \frac{1}{n}\,\mathrm{D}_0^2\right)\sigma^2 + \sigma^2 \\
&\qquad (\because \text{式 (3)}\ \ \mathrm{V}(Y_0) = \left\{\frac{1}{n} + \frac{1}{n}\sum_{j=1}^{p}\sum_{l=1}^{p}(x_{j0} - \bar{x}_j)(x_{l0} - \bar{x}_l)s^{jl}\right\}\sigma^2 \text{ より}) \\
&= \left(1 + \frac{1}{n} + \frac{1}{n}\,\mathrm{D}_0^2\right)\sigma^2
\end{aligned} \tag{10}$$

となる. これを用いて, $Y_0 - y$ を標準化し, σ^2 を不偏推定値 $\mathrm{V}_e = \frac{\mathrm{F}(\hat{a}_0, \hat{a}_1, \cdots, \hat{a}_p)}{n-p-1}$ で置き換えると, 式 (11) を得る. t' は自由度 $n-p-1$ の t 分布に従う.

$$t' = \frac{Y_0 - y}{\sqrt{\left(1 + \frac{1}{n} + \frac{1}{n}\,\mathrm{D}_0^2\right)\mathrm{V}_e}} \tag{11}$$

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,

$$-t_\alpha(n-p-1) \leq \frac{Y_0 - y}{\sqrt{\left(1 + \frac{1}{n} + \frac{1}{n}\,\mathrm{D}_0^2\right)\mathrm{V}_e}} \leq t_\alpha(n-p-1)$$

すなわち,

$$Y_0 - t_\alpha(n-p-1)\sqrt{\left(1 + \frac{1}{n} + \frac{1}{n}\,\mathrm{D}_0^2\right)\mathrm{V}_e} \leq y \leq Y_0 + t_\alpha(n-p-1)\sqrt{\left(1 + \frac{1}{n} + \frac{1}{n}\,\mathrm{D}_0^2\right)\mathrm{V}_e} \tag{12}$$

のように与えられる.

2 §1.13 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回帰係数や予測値を計算してきた. しかし, 実際に現象を分析する場合には, 目的変数 y に影響を及ぼすかもしれないと考え

られる変数は多数あり, その中から次のようなことを考慮し, しかも実質科学的にも重要な変数を取り上げるのが普通である.

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, 回帰係数の推定値 $\hat{a}_j$, 目的変数の予測値 Y_0 は不偏であるが, 誤差分散の推定値 V_e の自由度 $n-p-1$ が小さくなり, $\hat{a}_i$ や Y_0 の推定精度が悪くなる.
2. 必要な変数 (真の回帰係数が 0 でない変数) が回帰モデルの中から漏れている場合, 回帰係数の推定値, 目的変数の予測値は偏りをもち, また誤差分散の推定値 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には, 分散共分散行列 $V=(s_{jl})$ 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は悪くなる. 特に説明変数の中のひとつと残りの変数との重相関係数が $\mathrm{R}=1$ のときには分散共分散行列 (s_{jl}) の行列式は 0 になり逆行列が存在しないため, 回帰係数の推定値 $\hat{a}_j$ は得られない. このような場合多重共線性の問題があると言う.

説明変数の候補の中から最良な変数を選択して回帰式を求めるための統計的方法として次のような方法が提案されている.

(1) 総あたり法

p 個の説明変数の候補の中から $1\sim p$ 個の変数の可能なすべての部分集合に対応する 2^p-1 通りの回帰モデルを検討する方法. p が大きくなると場合の数は急速に大きくなり, 計算時間が膨大になる.

(2) 前進選択法

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

1. 目的変数 y との単相関が最大 (すなわち, 1 つずつ順番に変数を採用してみて回帰式を計算したとき, 回帰係数の検定のための t の絶対値または F 値が最大) の変数を選び, 回帰係数がゼロであるという仮説の検定をして仮説が棄却されなければどの変数も回帰モデルに含めない. 仮説が棄却されればこの変数を取り込んで次のステップに進む.
2. 既に入っている変数に加えて残りの変数を 1 つずつ順番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 であると言う仮説の検定をおこない, 仮説が棄却されなければ終了. 仮説が棄却されれば選ばれた変数を取り込んで次のステップへ進む.
3. 回帰式を計算する. もしモデルに全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 2 に戻る

(3) 後退消去法

説明変数の候補すべてが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.

1. モデルに含まれている変数の各々に対する回帰係数検定のための t または F 値を計算し, その中の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へ進む. 棄却されれば終了.
2. 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直してステップ 1 に戻る.

(4) 逐次法

前進選択法では 1 度入った変数は落とされることがないという点を改良して, 次のような手順で変数を

増減させる.

1. 目的変数 y 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おこない, 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取り込んで次のステップに進む.
2. 既に入っている変数に加えて残りの変数を 1 つずつ順番に採用してみて偏相関係数が最大の変数を選ぶ. 回帰係数が 0 であるという仮説が棄却されなければ終了. 棄却されれば選ばれた変数を取り込んで次のステップに進む.
3. 回帰式を計算して各変数について回帰係数の検定をおこない, 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おとす.
4. すべての変数が取り込まれていれば終了. そうでなければステップ 2 に戻る.

上記 4 つの方法において, 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 ♣ 式 (19) より,

$$|t| = \frac{|\hat{a}_j - a_j^{(0)}|}{\sqrt{\frac{s^{jj}\,\mathrm{V}_e}{n}}} \geq t_\alpha(n-p-1)$$

で, $a_j^{(0)} = 0$ とおき,

$$t = \frac{\hat{a}_j}{\sqrt{s^{jj}\,\mathrm{V}_e\,/n}} \tag{13}$$

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, $|t| \leq t_\alpha(n-p-1)$ ならば仮説を採択する. また, 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,

$$\mathrm{F} = \frac{\hat{a}_j^2}{s^{jj}\,\mathrm{V}_e\,/n} \tag{14}$$

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, $|\,\mathrm{F}\,| \geq \mathrm{F}^1_{n-p-1}(\alpha)$ ならば仮説を棄却することで同じ結果を得ることができる.

付録 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 より,

$$\hat{a}_1 = \frac{\begin{vmatrix} s_{y1} & s_{12} \\ s_{y1} & s_{22} \end{vmatrix}}{\begin{vmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{vmatrix}} \tag{15}$$

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

$$\begin{aligned} |\boldsymbol{V}| &= s_{11}s_{22} - s_{12}s_{21} \\ &= s_{11}s_{22}\left(1 - \frac{s_{12}}{\sqrt{s_{11}s_{22}}}\frac{s_{21}}{\sqrt{s_{11}s_{22}}}\right) \\ &= s_{11}s_{22}(1-r^2) \end{aligned} \tag{16}$$

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 ± 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 (15) の分子の行列式 $(s_{y1}s_{22} - s_{12}s_{y1})$ の変化に大きく影響されることがわかる. したがって, 変数の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落ちる. 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 (15) は発散してしまし, 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.