# 多変量解析　～重回帰分析～

2006年4月21日(金)　南　慶典

# 重回帰分析とは？

重回帰分析とは

- 複数の説明変数から目的変数との関係性を予測、評価
- 説明変数（数量データ）は目的変数を説明するのに有効であるか
- 得られた関係性より未知のデータの妥当性を判断する。これを重回帰分析という。

つまり、どんなことをするのか？

① 最小2乗法により重回帰モデルを想定
② 自由度調整済寄与率を求め、得られた回帰式の性能を評価する
③ 説明変数の選択（変数選択）を行い、有用な変数を選択する
④ 残差とテコ比の検討を行い、得られた回帰式の妥当性を検討
⑤ 将来得られるデータ値を予測

# 重回帰分析の具体例

表は東京のある駅の徒歩圏内の中古マンションに関するデータである。

| サンプルNO. | 広さ　x1 | 築年数　x2 | 価格　y |
|---|---|---|---|
| 1 | 51 | 16 | 3.0 |
| 2 | 38 | 4 | 3.2 |
| 3 | 57 | 16 | 3.3 |
| 4 | 51 | 11 | 3.9 |
| 5 | 53 | 4 | 4.4 |
| 6 | 77 | 22 | 4.5 |
| 7 | 63 | 5 | 4.5 |
| 8 | 69 | 5 | 5.4 |
| 9 | 72 | 2 | 5.4 |
| 10 | 73 | 1 | 6.0 |

このデータに基づいて知りたいことは次の通りである。

（1）価格は広さと築年数とによって予測できるだろうか。
（2）予測できるとすればその精度はどのくらいか。
（3）同じ地区で$x_1=70, x_2=10, y=5.8$を掲示された。価格は妥当か。

などを重回帰分析で検討する。

# 説明変数が2個の場合の解析方法

## 最小2乗法による回帰式の推定

表1のデータに関して次の重回帰モデル(回帰モデル)を想定する。

$$y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\varepsilon_i \quad \varepsilon_i \sim N(0,\sigma^2)$$

i番目の予測値　$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\hat{\beta}_2 x_{i2}$

i番目の残差　$e_i=y_i-\hat{y}_i$

**最小2乗法**

実測値と予測値の残差平方和を最小にする $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p$ を求める方法

残差平方和 $S_e$

$$S_e=\sum_{i=1}^{n} e_i^2=\sum_{i=1}^{n}\left\{y_i-\left(\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\hat{\beta}_2 x_{i2}\right)\right\}^2$$

これを最小にする $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ を求める

# 正規方程式

$S_e$ を $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ のそれぞれで偏微分して 0 とおくと

$$\frac{\partial S_e}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) = 0 \quad \cdots (1)$$

$$\frac{\partial S_e}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n} x_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) = 0 \quad \cdots (2)$$

$$\frac{\partial S_e}{\partial \hat{\beta}_2} = -2\sum_{i=1}^{n} x_{i2}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) = 0 \quad \cdots (3)$$

（1）、（2）、（3）式を整理すれば、

$$n\hat{\beta}_0 + \hat{\beta}_1 \Sigma x_{i1} + \hat{\beta}_2 \Sigma x_{i2} = \Sigma y_i$$

$$\hat{\beta}_0 \Sigma x_{i1} + \hat{\beta}_1 \Sigma x_{i1}^2 + \hat{\beta}_2 \Sigma x_{i1} x_{i2} = \Sigma x_{i1} y_i$$

$$\hat{\beta}_0 \Sigma x_{i2} + \hat{\beta}_1 \Sigma x_{i1} x_{i2} + \hat{\beta}_2 \Sigma x_{i2}^2 = \Sigma x_{i2} y_i$$

$$\hat{\beta}_0 = \frac{\Sigma y_i}{n} - \hat{\beta}_1 \frac{\Sigma x_{i1}}{n} - \hat{\beta}_2 \frac{\Sigma x_{i2}}{n}$$

と変形して、下の2つに代入していく。

これらは、$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ に関する連立方程式であり、**正規方程式**と呼ぶ。

# 偏回帰係数を求める

正規方程式を変形して、以下の連立方程式を得る。

$$\hat{\beta}_1 S_{11} + \hat{\beta}_2 S_{12} = S_{1y}$$
$$\hat{\beta}_1 S_{12} + \hat{\beta}_2 S_{22} = S_{2y}$$

ただし、各変数の平方和と偏差積和を次のように定義する。

$$S_{11} = \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2 \qquad S_{22} = \sum_{i=1}^{n} (x_{i2} - \bar{x}_2)^2 \qquad S_{12} = \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad S_{1y} = \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(y_i - \bar{y}) \qquad S_{2y} = \sum_{i=1}^{n} (x_{i2} - \bar{x}_2)(y_i - \bar{y})$$

行列で表現すると、

$$\begin{bmatrix} S_{11} & S_{12} \\ S_{12} & S_{22} \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} S_{1y} \\ S_{2y} \end{bmatrix}$$

この連立方程式を解けば、偏回帰係数 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ が求まる。

# 多重共線性に注意する（1）

多重共線性が存在するとは･･･

　　行列の逆行列が存在しない状況をいう。

実際に、偏回帰係数の解を求めていくと

$$\begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix} = \begin{bmatrix}S_{11} & S_{12}\\ S_{12} & S_{22}\end{bmatrix}^{-1}\begin{bmatrix}S_{1y}\\ S_{2y}\end{bmatrix}$$

$$= \frac{1}{S_{11}S_{22}-S_{12}^2}\begin{bmatrix}S_{22}S_{1y}-S_{12}S_{2y}\\ -S_{12}S_{1y}+S_{11}S_{2y}\end{bmatrix}$$

すなわち $S_{11}S_{22}-S_{12}^2$ が0であると、解は無数に存在するか、まったく存在しない

例）　$S_{11}=1,\ S_{22}=4,\ S_{12}=2$ である場合

$$\hat{\beta}_1+2\hat{\beta}_2 = S_{1y}$$
$$2\hat{\beta}_1+4\hat{\beta}_2 = S_{2y}$$

# 多重共線性に注意する(2)

$$S_{11}S_{22}-S_{12}^2 = 0 \Leftrightarrow \frac{S_{12}^2}{(S_{11}S_{22})} = 1$$

$$\Leftrightarrow r_{x_1x_2}^2 = \left\{\frac{S_{12}}{\sqrt{S_{11}S_{22}}}\right\}^2 = 1$$

$$\Leftrightarrow r_{x_1x_2} = \pm 1$$

すなわち、$x_1$ と $x_2$ の相関係数が1または－1の時に多重共線性が存在する。

相関係数が±1となるのは点 $(x_1, x_{i2})\ (i=1,2,\cdots,n)$ の全てが1直線上に並んでいる場合であり、$x_1$ と $x_2$ が共通の直線状にある(共線)。この場合、他方の情報は不要である。

「予測を行う」という観点から偏回帰係数が定められるので，重回帰式に含まれた変数相互間の関連で符号が決められるためである。

多重共線性の存在の有無を考慮することが必要

# 寄与率と自由度調整済み寄与率（1）

～回帰式の評価～

まず、残差平方和を整理しておく

$$S_e = \sum_{i=1}^{n} \left\{ y_i - \left( \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} \right) \right\}^2$$

$$= \sum_{i=1}^{n} \left\{ y_i - \bar{y} - \hat{\beta}_1 ( x_{i1} - \bar{x_1} ) x_{i1} - \hat{\beta}_2 ( x_{i2} - \bar{x_2} ) \right\}^2$$

$$= S_{yy} + \hat{\beta_1^2} S_{11} + \hat{\beta_2^2} S_{22} - 2 \hat{\beta}_1 S_{1y} - 2 \hat{\beta}_2 S_{2y} + 2 \hat{\beta}_1 \hat{\beta}_2 S_{12}$$

$$= S_{yy} + \hat{\beta}_1 ( \hat{\beta}_1 S_{11} + \hat{\beta}_2 S_{12} ) + \hat{\beta}_2 ( \hat{\beta}_1 S_{12} + \hat{\beta}_2 S_{22} ) - 2 \hat{\beta}_1 S_{1y} - 2 \hat{\beta}_2 S_{2y}$$

$$= S_{yy} - ( \hat{\beta}_1 S_{1y} + \hat{\beta}_2 S_{2y} )$$

重回帰モデルの誤差 $\varepsilon$ の母分散 $\sigma^2$ を次のように推定することができる

$$\hat{\sigma^2} = V_e = \frac{S_e}{\phi_e} = \frac{S_e}{n-3}$$

# ちょっとした準備　～平方和の分解～

偏差平方和　＝　残差平方和　＋　回帰平方和

$$
\begin{aligned}
S_{yy} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\
&= \Sigma \left\{ y_i - \left( \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} \right) + \left( \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} \right) - \bar{y} \right\}^2 \\
&= \Sigma \left\{ y_i - \left( \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} \right) \right\}^2 + \Sigma \left\{ \left( \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} \right) - \bar{y} \right\}^2 \\
&\qquad + 2\Sigma \left\{ y_i - \left( \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} \right) \right\} \left\{ \left( \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} \right) - \bar{y} \right\} \\
&= \Sigma \left\{ y_i - \left( \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} \right) \right\}^2 + \Sigma \left\{ \left( \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} \right) - \bar{y} \right\}^2 \quad \cdots \ (4)
\end{aligned}
$$

# 補足スライド～ここはいんじゃない？～

$$\sum_{i=1}^{n}\left\{y_i-\left(\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\hat{\beta}_2 x_{i2}\right)\right\}\left\{\left(\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\hat{\beta}_2 x_{i2}\right)-\bar{y}\right\}$$

$$= \sum_{i=1}^{n} e_i\left\{\left(\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\hat{\beta}_2 x_{i2}\right)-\bar{y}\right\}$$

$$= (\hat{\beta}_0-\bar{y})\Sigma e_i+\hat{\beta}_1\Sigma x_{i1}e_i+\hat{\beta}_2\Sigma x_{i2}e_i = 0$$

$$S_{12} = \Sigma(x_{i1}-\bar{x}_1)(e_i-\bar{e})$$

$$= \Sigma x_{i1}e_i-\bar{x}_1\Sigma e_i = 0$$

$$\Sigma e_i = 0 \quad (\bar{e} = 0)$$

# 寄与率と自由度調整済寄与率（2）

$$S_R = \hat{\beta}_1 S_{1y} + \hat{\beta}_2 S_{2y}$$ とおくと

$$S_{yy} = \hat{\beta}_1 S_{1y} + \hat{\beta}_2 S_{2y} + S_e = S_R + S_e \quad \cdots \text{（5）}$$

（4）式と（5）式を見比べることにより、回帰平方和は

$$S_R = \Sigma \left\{ \left( \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} \right) - \bar{y} \right\}^2 = \hat{\beta}_1 S_{1y} + \hat{\beta}_2 S_{2y}$$

各平方和には、各自由度が対応している

$$S_{yy} \quad \phi_T = n - 1$$

$$S_R \quad \phi_R = 2$$

$$S_e \quad \phi_e = n - 3$$

# 寄与率と自由度調整済寄与率（3）

実測値と理論値の相関係数

$$R = \frac{\Sigma(y_i-\bar{y})(\hat{y}_i-\bar{y})}{\sqrt{\Sigma(y_i-\bar{y})^2\Sigma(\hat{y}_i-\bar{y})^2}}$$

これを**重相関係数**という。実測値と理論値がどれだけ一致しているかを求めている。

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

これを**寄与率（または決定係数）**と呼ぶ。これはyの変動のうちの回帰による変動の割合を表している。決定係数が大きければよいわけではなく、どんなに意味のない説明変数を加えてもこの値は上昇してしまう。よって自由度を用いて調整して、

$$R^{*2} = 1-\frac{S_e/\phi_e}{S_{yy}/\phi_T}$$

このような決定係数を**自由度調整済寄与率**という。

# 説明変数の選択（変数選択）

できるだけ説明変数は目的変数に効いている説明変数だけをモデルに含めたい。

意味のない説明変数を加えていっても、多重共線性の原因を招いてしまったり、
分析の精度が低くなる。

→ 説明変数の選択が重要になる。

**説明変数の選択基準**

- 目的変数と相関の高い変数を説明変数にする
- 単相関係数を用い、相関が0.7以上のものを説明変数にするのが一般的

説明変数の相互間で相関係数が1に近い場合、どちらか解釈しやすいものを選択する

$R^{*2}$ が増加する限り，追加された独立変数は有効であることを意味する。

# 変数の選択方法

変数減少法　・・・　すべての変数を取り込んだ段階から不要な変数を消去していく

変数増加法　・・・　定数項だけのモデルから有用な変数を追加していく

変数増減法　・・・　それら両方を取り入れた方法

・変数増加法

yの平方和 $S_{yy}$（自由度 $\phi_T = n-1$）と残差平方和 $S_e$（自由度 $\phi_e = n-p-1$）を用いて、不偏分散比の大きさを目安にしていく。

不偏分散比が、有意水準α（一般的には0.01or0.05）におけるF分布の値より大きければ有意。この値が大きい方の説明変数をモデルに取り込む。

$$F_0 = \frac{\left(S_{yy} - S_{e_{(MI)}}\right) / \left(\phi_T - \phi_{e_{(MI)}}\right)}{S_{yy} / \phi_T}$$

寄与率、自由度調整済寄与率を求め、そのモデルが妥当か評価する。

# 変数増加法による変数選択

次に変数を追加するかどうかは、残差平方和（M1）と加えたときの残差平方和（M2）の不偏分散比のF値を比較する。

$$F_0 = \frac{\left(S_{e_{(M1)}} - S_{e_{(M2)}}\right) / \left(\phi_{e_{(M1)}} - \phi_{e_{(M2)}}\right)}{S_{e_{(M2)}} / \phi_{e_{(M2)}}}$$

F値が大きければ（2以上）であれば、その変数を取り込む

式の分子はMODEL1からMODEL2に変更することにより残差平方和がどれくらい減少するのかを測る量を示している。

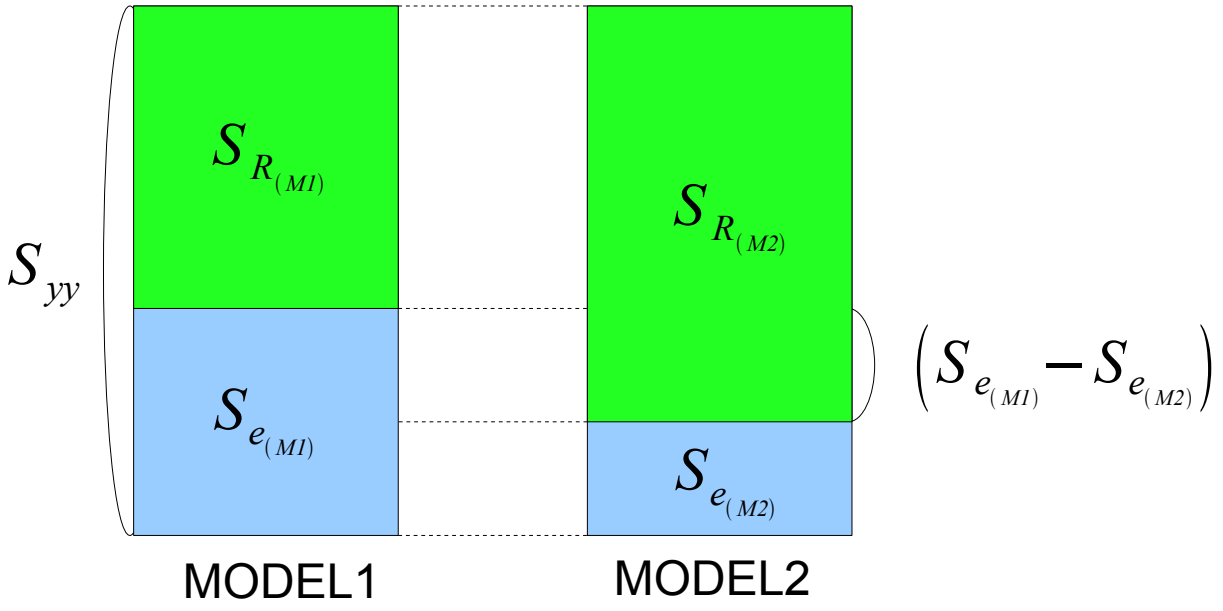
例） 変数が2つの場合、

MODEL1

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$$

MODEL2

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$



# 残差とテコ比の検討

～重回帰式の妥当性の評価～

残差 $e_k$ を標準化したもの、また残差のt値について求める

$$e'_k = \frac{e_k}{\sqrt{V_e}} \qquad t_k = \frac{e_k}{\sqrt{(1-h_{kk})V_e}}$$

・各値の絶対値が3.0以上または2.5以上である場合はサンプルが異常でないか検討

・できれば各説明変数を横軸にとり、標準化残差またtを縦軸にとって散布図を描く

→曲線的な傾向や、説明変数が大きくなるに従って残差のばらつきが系統的に変化していないかなどを検討する

# テコ比？

予測値の第kサンプルを表すテコ比を使って表すと、

$$\hat{y}_k = h_{k1} y_1 + h_{k2} y_2 + \cdots + h_{kk} y_k + \cdots + h_{kn} y_n$$

$y_k$ の係数 $h_{kk}$ をテコ比(レベレッジ)と呼び、

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

但し、$D_k^2$ をマハラノビスの距離の２乗と呼ばれ、判別分析で重要な役割を果たす

$$D_k^2 = (n-1)\left\{(x_{k1}-\bar{x}_1)^2 S^{11} + 2(x_{k1}-\bar{x}_1)(x_{k2}-\bar{x}_2) S^{12} + (x_{k2}-\bar{x}_2)^2 S^{22}\right\}$$

但し、$\begin{bmatrix} S^{11} & S^{12} \\ S^{12} & S^{22} \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} \\ S_{12} & S_{22} \end{bmatrix}^{-1}$

このテコ比が大きすぎると、$\hat{y}_k$ の値が $y_k$ の値の変動によって強く影響を受けるので望ましくない

**2.5{（説明変数の個数）＋1}／n ＝ 2.5×（テコ比の平均）**

データ取得時に調整できるなら、テコ比がこれより小さくなるように工夫する

# 得られた回帰式の利用

回帰式の $\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2$ 推定量の確率分布

$$\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2 \quad \sim \quad N\left(\beta_0+\beta_1 x_1+\beta_2 x_{2,}\left\{\frac{1}{n}+\frac{D^2}{n-1}\right\}\sigma^2\right)$$

これを用いて、$x_{i1}$と $x_{i2}$を任意の値 $x_{01}$, $x_{02}$ に設定して、母回帰の区間推定や予測区間を構成することができる。

母回帰の信頼率95%の信頼区間は次のように構成する。

$$\hat{\beta}_0+\hat{\beta}_1 x_{01}+\hat{\beta}_2 x_{02} \pm t(\phi_e,\ 0.05)\sqrt{\left\{\frac{1}{n}+\frac{D_0^2}{n-1}\right\}V_e}$$

と設定した場合に回帰直線上の縦座標の信頼区間である。
信頼率95%の予測区間は次のように計算する。

$$\hat{\beta}_0+\hat{\beta}_1 x_{01}+\hat{\beta}_2 x_{02} \pm t(\phi_e,\ 0.05)\sqrt{\left\{1+\frac{1}{n}+\frac{D_0^2}{n-1}\right\}V_e}$$