回帰分析

回帰分析は、ある変数（被説明変数）が他の幾つかの変数（説明変数）で構成される関数によって決まる関係を分析する手法の一つであり、これの目的は説明変数と被説明変数との定量的な関係の構造（モデルと呼ぶことがある）を求めることである。活用の対象は自然科学や人文・社会科学など多岐に亘り、構造式が線形関数の場合線形回帰、それ以外の場合非線形回帰と呼ぶ。ここでは線形回帰のみを扱う。例題は最後の節に示している。

線形回帰モデル

人口の多い都市ほど交通発生量は多いであろう。この関係を交通量Y、人口 X とすると都市という母集団において$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (i = 1,2,\cdots,n)$が成立すると考える。$\varepsilon_i$は誤差項である。今、誤差項$\varepsilon_i$を無視して$Y_i$の予測値$\hat{Y}_i = \beta_0 + \beta_1 X_i$を考えよう。未知となっているパラメータ$\beta_0, \beta_1$が一定の計算作業により$\hat{\beta}_0, \hat{\beta}_1$と推計されたとして、これを固定して用いると標本観測値の予測値$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$が求まる。但し観測値$y_i$と$\hat{y}_i$が一致するとは限らない。では、$\hat{\beta}_0, \hat{\beta}_1$をどのように求めたらよいだろうか。

先の$\hat{Y}_i = \beta_0 + \beta_1 X_i$において$X_i$には標本観測値を用いて$\hat{Y}_i = \beta_0 + \beta_1 x_i$というモデルを考えることにより、標本毎の予測値$\hat{y}_i = \beta_0 + \beta_1 x_i$を得る。

これの平均二乗誤差$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / n = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 / n$を最小にするような$\beta_0, \beta_1$を$\hat{\beta}_0, \hat{\beta}_1$として固定して用いればよい。実際に求めてみよう。パラメータβ_0, β_1の関数について最小化問題なので、平均二乗誤差を偏微分して0になるものを求めことになり、解くと

$\frac{\partial S}{\partial \hat{\beta}_0} = 0, \quad \frac{\partial S}{\partial \hat{\beta}_1} = 0$ より

正規方程式$\begin{cases} \hat{\beta}_0 n + \hat{\beta}_1 \sum_i x_i = \sum_i y_i \\ \hat{\beta}_0 \sum_i x_i + \hat{\beta}_1 \sum_i x_i^2 = \sum_i x_i y_i \end{cases}$を得る。これを整理して、

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \dfrac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \end{cases}$$ となる。

ここで得られた$\hat{\beta}_0, \hat{\beta}_1$を用いて、再度最小二乗誤差をみてみると、

$$S^* = \frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$$
$$= \frac{1}{n}\sum_i (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 = \frac{1}{n}\sum_i (y_i - (\bar{y} - \hat{\beta}_1 \bar{x}_i + \hat{\beta}_1 x_i))^2 = \frac{1}{n}\sum_i ((y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}_i))^2$$
$$= \frac{1}{n}\sum_i (y_i - \bar{y})^2 - 2\hat{\beta}_1 \left\{ \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y}) \right\} + {\hat{\beta}_1}^2 \left\{ \frac{1}{n}\sum_i (x_i - \bar{x})^2 \right\}$$
$$= S_{yy} - 2\frac{S_{xy}^2}{S_{xx}} + \frac{S_{xy}^2}{S_{xx}} = S_{yy}\left(1 - \frac{S_{xy}^2}{S_{xx} S_{yy}}\right)$$

となる。なお、$S_{yy} \equiv \frac{1}{n}\sum_i (y_i - \bar{y})^2$,　$S_{xy} \equiv \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})$,　$S_{xx} \equiv \frac{1}{n}\sum_i (x_i - \bar{x})^2$ である。

よって、$0 \le S^* \le S_{yy}$,　$0 \le \frac{S^*}{S_{yy}} \le 1$　であり、

$\frac{S_{xy}^2}{S_{xx} S_{yy}} \equiv r^2$,　$r^2 = 1 - \frac{S^*}{S_{yy}}$,　$\therefore 1 \ge r^2 \ge 0 \rightarrow -1 \le r \le 1$　のようになる。

rを相関係数、r^2を寄与率と呼ぶ。

平均二乗誤差を最小にするようにパラメータ値を決める方法を最小二乗法という。なお平均二乗誤差のようにサンプル数(n)で除すことなく、誤差の二乗和でも結果は同じになる。

これまでは説明変数が１個であったが、一般化して２個以上を考える。X_{ij}は確定値である。

$$\hat{Y}_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$$

これの標本毎の予測値は次のようになる。

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

誤差の二乗和を考え、これの最小値を与えるパラメータベクトルを求める。

$$S=\sum_i (y_i-\hat{y}_i)^2=\sum_i (y_i-\beta_0-\beta_1 x_{i2}-\cdots-\hat{\beta}_k x_{i,k})^2 \to \min,$$

微分すると、

$$\frac{\partial S}{\partial \beta_0}=0,\quad 2\sum_i (y_i-\hat{\beta}_0-\hat{\beta}_1 x_{i1}-\cdots-\hat{\beta}_k x_{i,k})(-1)=0$$
$$\frac{\partial S}{\partial \beta_1}=0,\quad 2\sum_i (y_i-\hat{\beta}_0-\hat{\beta}_1 x_{i1}-\cdots-\hat{\beta}_k x_{i,k})(-x_{i1})=0$$
$$\frac{\partial S}{\partial \beta_2}=0,\quad 2\sum_i (y_i-\hat{\beta}_0-\hat{\beta}_2 x_{i1}-\cdots-\hat{\beta}_k x_{i,k})(-x_{i2})=0$$
……
$$\frac{\partial S}{\partial \beta_k}=0,\quad 2\sum_i (y_i-\hat{\beta}_1-\hat{\beta}_2 x_{i1}-\cdots-\hat{\beta}_k x_{i,k})(-x_{i,k})=0$$

となる。なお、ここで

$$\sum_i (y_i-\hat{\beta}_0 x_{i1}-\hat{\beta}_1 x_{i1}-\cdots-\hat{\beta}_k x_{i,k})=\sum_i (y_i-\hat{y}_i)=\sum_i \hat{\varepsilon}_i=0$$
$$\sum_i (y_i-\hat{\beta}_0 x_{i1}-\hat{\beta}_1 x_{i1}-\cdots-\hat{\beta}_k x_{i,k})x_{ij}=\sum_i (y_i-\hat{y}_i)x_{ij}=\sum_i \hat{\varepsilon}_i x_{ij}=0,\quad j=1,\cdots,k$$

となっていることに注意しよう。ε_i の予測値（推計誤差）$\hat{\varepsilon}_i$ の平均値は 0 であり、$\hat{\varepsilon}_i$ と X_i は直交していることを示している。これは最小二乗法それ自体の特徴であり、母集団に関係なく常に成立する重要な性質である。

これを整理して、正規方程式は

$$\hat{\beta}_0\sum_i 1+\hat{\beta}_1\sum_i x_{i1}+\hat{\beta}_2\sum_i x_{i2}+\cdots+\hat{\beta}_k\sum_i x_{i,k}=\sum_i y_i$$
$$\hat{\beta}_0\sum_i 1+\hat{\beta}_1\sum_i x_{i1}^2+\hat{\beta}_3\sum_i x_{i1}x_{i2}+\cdots+\hat{\beta}_k\sum_i x_{i1}x_{i,k}=\sum_i x_{i1}y_i$$
$$\hat{\beta}_0\sum_i 1+\hat{\beta}_1\sum_i x_{i1}x_{i2}+\hat{\beta}_3\sum_i x_{i2}^2+\cdots+\hat{\beta}_k\sum_i x_{i2}x_{i,k}=\sum_i x_{i2}y_i$$
……
$$\hat{\beta}_0\sum_i 1+\hat{\beta}_1\sum_i x_{i1}x_{i,k}+\hat{\beta}_3\sum_i x_{i2}x_{i,k}+\cdots+\hat{\beta}_k\sum_i x_{i,k}^2=\sum_i x_{i,k}y_i$$

なお、第 1 式は $\hat{\beta}_1=\bar{y}-\hat{\beta}_2\bar{x}_2-\cdots-\hat{\beta}_k\bar{x}_k$ のように変形され、これを第 2 式以降に代入して整理すると、残る式群においてｊ番目式は、$\sum_{h=1}^{k}\beta_h a_{jh}=a_{jy},\ j=1,\cdots,k$ と簡潔に示される。

但 し $a_{lm} \equiv \sum_i (x_{il} - \bar{x}_l)(x_{im} - \bar{x}_m), a_{ly} \equiv \sum_i (x_{il} - \bar{x}_l)(y_i - \bar{y}), \quad (l, m = 2, \cdots, k); a_{lm} = a_{ml})$

である。
先のΣ記号による式を行列表現すると、

$$\tilde{\hat{\boldsymbol{\beta}}}^t = \begin{bmatrix} \beta_1 & \beta_2 & \cdots & \beta_k \end{bmatrix}, A_y^t = \begin{bmatrix} a_{1y} & a_{2y} & \cdots & a_{ky} \end{bmatrix}, A = \begin{bmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{bmatrix}, \cdots A^{-1} = \begin{bmatrix} a^{11} & \cdots & a^{1k} \\ \vdots & \ddots & \vdots \\ a^{k1} & \cdots & a^{kk} \end{bmatrix}$$

$\mathbf{A}\tilde{\hat{\boldsymbol{\beta}}} = \mathbf{A}_\mathbf{y}$ であるから、$\tilde{\hat{\boldsymbol{\beta}}} = \mathbf{A}^{-1}\mathbf{A}_\mathbf{y}$ となる。

行列表現は見通しがよいので、一括して行列表現で解こう。このときパラメータは定数項β_0がついてきて説明変数の個数よりも一つ多いことになるので、この定数項に１に確定した変数を与え、パラメータと説明変数の数を揃えて、次の見通しの良い式を考える。

$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}$$

$$\mathbf{y}^t = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}, \quad \hat{\mathbf{y}}^t = \begin{bmatrix} \hat{y}_1 & \hat{y}_2 & \cdots & \hat{y}_n \end{bmatrix}, \quad \boldsymbol{\beta}^t = \begin{bmatrix} \beta_1 & \beta_2 & \cdots & \beta_k \end{bmatrix}$$

但し、 $\mathbf{X} = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \cdots & x_{nk} \end{bmatrix}$

$\mathbf{X^tX}$ は正則であるとする。

$$S = (\mathbf{y} - \hat{\mathbf{y}})^t (\mathbf{y} - \hat{\mathbf{y}}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad \to \min$$

$$S = \mathbf{y^t y} - \mathbf{y^t X}\boldsymbol{\beta} - \boldsymbol{\beta^t X^t y} + \boldsymbol{\beta^t X^t X}\boldsymbol{\beta} = \mathbf{y^t y} - 2\mathbf{y^t X}\boldsymbol{\beta} + \boldsymbol{\beta^t X^t X}\boldsymbol{\beta}$$

$$\frac{\partial S}{\partial \boldsymbol{\beta}} = \mathbf{0}$$

$$-2\mathbf{y^t X} + 2\hat{\boldsymbol{\beta}}^\mathbf{t}\mathbf{X^t X} = \mathbf{0}, \quad \mathbf{X^t X}\hat{\boldsymbol{\beta}} = \mathbf{X^t y}$$

$$\therefore \hat{\boldsymbol{\beta}} = (\mathbf{X^t X})^{-1} \quad \mathrm{X}^\mathbf{t}\mathbf{y}$$

exerxise;

$$\mathbf{P} = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix}, \mathbf{A} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

$$\mathbf{P^tAP} = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix}^t \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = p_1^2 + 4p_1p_2 + p_2^2$$

$$\frac{\partial(\mathbf{P^tAP})}{\partial \mathbf{P}} = \begin{bmatrix} \dfrac{\partial(P^tAP)}{\partial p_1} \\ \dfrac{\partial(P^tAP)}{\partial p_2} \end{bmatrix} = \begin{bmatrix} 2p_1 + 4p_2 \\ 4p_1 + 2p_2 \end{bmatrix} = 2\begin{bmatrix} p_1 + 2p_2 \\ 2p_1 + p2 \end{bmatrix} = 2\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = 2\mathbf{AP}$$

２．モデルの性質

前章では誤差項ε_iの性格については何も仮定を置かず話を進めた。回帰式を求めるだけなら、これで十分といえよう。しかし、モデルの推計精度や統計的性質を考慮することが必要となる場合が多い。そこで誤差項ε_iが確率的振る舞いをもつものとしてモデルを考えよう。

$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$

この式は、誤差項ε_iを標本観測値の推計値$\hat{y}_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$と標本観測値$y_i$とのずれを表している。

我々は$\hat{y}_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$のパラメータ$\beta_j$を$\hat{\beta}_j$として決める作業を行うのだから、$\hat{y}_i$値は確定項となり確率変数である誤差項$\varepsilon_i$の影響は$y_i$の方に現れるので、$y_i$を確率変数として捉え、$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$と書き直して確率モデルを考える。行列表現すると

$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

$\mathbf{Y}^t = [Y_1 \quad Y_2 \quad \cdots \quad Y_n]$, $\boldsymbol{\beta}^t = [\beta_1 \quad \beta_2 \quad \cdots \quad \beta_k]$ $\boldsymbol{\varepsilon}^t = [\varepsilon_1 \quad \varepsilon_2 \quad \cdots \quad \varepsilon_k]$

但し、$\mathbf{X} = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \cdots & x_{nk} \end{bmatrix}$

モデルの推計精度などをみるために誤差項ε_iに次の四つの仮定がなされる。つまり誤差は、平均が 0 であり分散が等しく、互いに独立で同一の分布に従うものと仮定している。

1)不偏性　$E(\varepsilon_i) = 0$

2)等分散性　$V(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$

3)独立性　$Cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_i) = 0 \quad (i \neq j)$

4)正規性　$N(0, \sigma^2)$

我々は標本値を扱うことで具体的なモデルを構築することになる。標本値を確率変数として扱うときに、標本変量と呼ぶことがある。標本変量の関数を統計量といい、よって統計量は確率変数であり、標本変量の実現値（標本値）から求められる統計量の実現値を統計値と呼んでいる。標本の全集合を母集団（無限母集団と有限母集団とがある）といい、これの確率分布及びそのパラメータ（平均値や分散など）は、母数と呼ばれ平均値や分散を母平均、母分散（得られた標本の平均と分散を標本平均、標本分散）という。得られた標本による母数の推定にあたり、点推定法において母数推定に用いられる統計量

$\Theta = f(X_1, X_2, \cdots, X_n)$ を推定量という。次の種類がある。

1)不偏性　$E(\Theta) = \theta$　不偏推定量

2)一致性　$\lim_{n\to\infty} P\{|\Theta - \theta| > \varepsilon\} = 0$ 一致推定量

3)有効性　$V(\Theta)$　min　有効推定量

4)十分性　推定量 Θ が母数 θ について標本のもつ情報をすべてもっている　十分推定量

最小二乗法によればパラメータの不偏性が保証されていることが知られている。

(1)偏回帰係数の推定量 $\hat{\beta}$ は、β の不偏推定量である。

まず、単回帰モデルの不偏パラメータを調べる。単回帰モデルは β_0 から重回帰モデルは β_1 から添え字を使っていることに注意。

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$$E(\hat{\beta}_1) = E\left(\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}\right) = \frac{E\left(\sum_i (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \varepsilon_i - \beta_0 - \beta_1 \bar{x})\right)}{\sum_i (x_i - \bar{x})^2}$$

$$= \frac{E\left(\sum_i (x_i - \bar{x})(\beta_1 (x_i - \bar{x}) + \varepsilon_i)\right)}{\sum_i (x_i - \bar{x})^2} = \frac{\beta_1 \sum_i (x_i - \bar{x})^2 + \sum_i (x_i - \bar{x}) E(\varepsilon_i)}{\sum_i (x_i - \bar{x})^2} = \beta_1 + \frac{\sum_i (x_i - \bar{x}) E(\varepsilon_i)}{\sum_i (x_i - \bar{x})^2} = \beta_1$$

$$E(\hat{\beta}_0) = E\left(\bar{y} - \hat{\beta}_1 \bar{x}\right) = E\left\{\beta_0 + \beta_1 \bar{x} + \frac{\sum_i \varepsilon_i}{n} - \left(\beta_1 + \frac{\sum_i (x_i - \bar{x})\varepsilon_i}{\sum_i (x_i - \bar{x})^2}\right)\bar{x}\right\}$$

$$= \beta_0 + \sum_i \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}\right) E(\varepsilon_i) = \beta_0$$

重回帰モデルで説明変数 2 個の場合の場合を考える。

$\hat{\beta}_2$ の期待値を求めてみよう。$\hat{\beta}_2 = a_{1y} a^{21} + a_{2y} a^{22}$ であったから

$$E(\hat{\beta}_2) = E\left((\beta_1 a_{11} + \beta_2 a_{12}) \frac{(-a_{12})}{a_{11}a_{22} - a_{12}a_{21}} + (\beta_1 a_{12} + \beta_2 a_{22}) \frac{a_{11}}{a_{11}a_{22} - a_{12}a_{21}}\right) = \beta_2$$ となる。

重回帰モデルの一般的な形についてみる。

$$E[\hat{\boldsymbol{\beta}}] = E[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}] = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t E[\mathbf{Y}] = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t E[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}] = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

(2) $\hat{\beta}$ の分散・共分散行列は Y_i が相互に無相関で共通の分散 σ^2 をもつならば $V\left[\hat{\boldsymbol{\beta}}\right]=\sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$

である。

単回帰モデルで確認すると β_1 は次のようになり、上記と一致する。

$$V(\hat{\beta}_1)=E\left(\hat{\beta}_1-E(\hat{\beta}_1)\right)^2=E(\hat{\beta}_1)^2-\beta_1^2$$

$$E(\hat{\beta}_1)^2=E\left(\frac{\sum_i(x_i-\bar{x})(y_i-\bar{y})}{\sum_i(x_i-\bar{x})^2}\right)^2=E\left(\frac{\sum_i(x_i-\bar{x})(\beta_1(x_i-\bar{x})+\varepsilon_i)}{\sum_i(x_i-\bar{x})^2}\right)^2$$

$$=E\left(\frac{\beta_1^2\left(\sum_i(x_i-\bar{x})^2\right)^2+2\beta_1\sum_i(x_i-\bar{x})^2\sum_i(x_i-\bar{x})\varepsilon_i+\left(\sum_i(x_i-\bar{x})\varepsilon_i\right)^2}{\left(\sum_i(x_i-\bar{x})^2\right)^2}\right)=\beta_1^2+E\left(\frac{\left(\sum_i(x_i-\bar{x})\varepsilon_i\right)^2}{\left(\sum_i(x_i-\bar{x})^2\right)^2}\right)$$

$$\therefore V(\hat{\beta}_1)=E\left(\frac{\left(\sum_i(x_i-\bar{x})\varepsilon_i\right)^2}{\left(\sum_i(x_i-\bar{x})^2\right)^2}\right)=E\left(\frac{\left(\sum_i x_i\varepsilon_i\right)^2-2\left(\sum_i x_i\varepsilon_i\sum_i\bar{x}\varepsilon_i\right)+\left(\sum_i\bar{x}\varepsilon_i\right)^2}{\left(\sum_i(x_i-\bar{x})^2\right)^2}\right)=E\left(\frac{\sum_i x_i^2\varepsilon_i^2-2\bar{x}\sum_i x_i\varepsilon_i^2+\bar{x}^2\sum_i\varepsilon_i^2}{\left(\sum_i(x_i-\bar{x})^2\right)^2}\right)$$

$$=E\left(\frac{\sum_i\varepsilon_i^2(x_i^2-2\bar{x}x_i+\bar{x}^2)}{\left(\sum_i(x_i-\bar{x})^2\right)^2}\right)=\frac{\sum_i E\left(\varepsilon_i^2\right)(x_i-\bar{x})^2}{\left(\sum_i(x_i-\bar{x})^2\right)^2}=\frac{\sigma^2\sum_i(x_i-\bar{x})^2}{\left(\sum_i(x_i-\bar{x})^2\right)^2}=\frac{\sigma^2}{\sum_i(x_i-\bar{x})^2}$$

β_0 についても下式のようなり、一致する。

$$V(\hat{\beta}_0)=E\left(\hat{\beta}_0-E(\hat{\beta}_0)\right)^2=E(\hat{\beta}_0^2)-\beta_0^2$$

$$E(\hat{\beta}_0^2)=E\left[\left\{\beta_0+\sum_i\left(\frac{1}{n}-\frac{\bar{x}(x_i-\bar{x})}{\sum_i(x_i-\bar{x})^2}\right)\varepsilon_i\right\}^2\right]=\beta_0^2+E\left[\left\{\sum_i\left(\frac{1}{n}-\frac{\bar{x}(x_i-\bar{x})}{\sum_i(x_i-\bar{x})^2}\right)\varepsilon_i\right\}^2\right]$$

$$\therefore V(\hat{\beta}_0)=E\left[\left\{\sum_i\left(\frac{1}{n}-\frac{\bar{x}(x_i-\bar{x})}{\sum_i(x_i-\bar{x})^2}\right)\varepsilon_i\right\}^2\right]=E\left[\frac{\left(\sum_i\varepsilon_i\right)^2}{n}-\frac{2\sum_i\varepsilon_i}{n}\frac{\sum_i\bar{x}(x_i-\bar{x})\varepsilon_i}{\sum_i(x_i-\bar{x})^2}+\frac{\left(\bar{x}\sum_i(x_i-\bar{x})\varepsilon_i\right)^2}{\left(\sum_i(x_i-\bar{x})^2\right)^2}\right]=$$

$$=\frac{\sigma^2}{n}-E\left(\frac{2\bar{x}\sum_i\varepsilon_i\sum_i(x_i\varepsilon_i-\bar{x}\varepsilon_i)}{n\sum_i(x_i-\bar{x})^2}\right)+E\left(\frac{\bar{x}^2\left(\sum_i(x_i\varepsilon_i-\bar{x}\varepsilon_i)\right)^2}{\left(\sum_i(x_i-\bar{x})^2\right)^2}\right)=\frac{\sigma^2}{n}-0+\frac{\bar{x}^2\sigma^2\sum_i(x_i-\bar{x})^2}{\left(\sum_i(x_i-\bar{x})^2\right)^2}=\frac{\sigma^2(\sum_i(x_i-\bar{x})^2+n\bar{x}^2)}{n\sum_i(x_i-\bar{x})^2}=\frac{\sigma^2\sum_i x_i^2}{n\sum_i(x_i-\bar{x})^2}$$

重回帰モデルの一般的な形についてみる。

$$V\left[\hat{\boldsymbol{\beta}}\right]=E\left[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^t\right]=*$$

$$\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}=(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}-(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}\boldsymbol{\beta}=(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})$$

$$*=E\left[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\cdot((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}))^t\right]=E\left[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\right]$$

$$=(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^tE\left[(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^t\right]\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}=**$$

$$E\left[(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^t\right]=E\left[\boldsymbol{\varepsilon}\cdot\boldsymbol{\varepsilon}^t\right]=\sigma^2\mathbf{I}$$

$$**=(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}=\sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$$

３．モデルの信頼性

得られたモデルがどの程度現象を再現できているのか気になるところである。そこで誤差についての特徴をみてみよう。

１）推計誤差の変動

推計誤差の変動は下式のように整理できる。

$$\sum_i (Y_i-\bar{Y})^2=\sum_i\left((Y_i-\hat{Y}_i)+(\hat{Y}_i-\bar{Y})\right)^2=\sum_i\left((Y_i-\hat{Y}_i)^2+2(Y_i-\hat{Y}_i)(\hat{Y}_i-\bar{Y})+(\hat{Y}_i-\bar{Y})^2\right)=*$$

$$\sum_i(Y_i-\hat{Y}_i)(\hat{Y}_i-\bar{Y})=\sum_i\hat{Y}_i(Y_i-\hat{Y}_i)-\bar{Y}\sum_i(Y_i-\hat{Y}_i)=\sum_{m=1}^{k}\beta_m\sum_i x_{im}(Y_i-\hat{Y}_i)-\bar{Y}\sum_i(Y_i-\hat{Y}_i)=\sum_{m=1}^{k}\beta_m\sum_i x_{im}\hat{\varepsilon}_i-\bar{Y}\sum_i\hat{\varepsilon}_i=0+0=0$$

$$*=\sum_i(Y_i-\hat{Y}_i)^2+\sum_i(\hat{Y}_i-\bar{Y})^2$$

つまり、$\sum_i(Y-\bar{Y})^2$（得られた観測データの偏差平方和、いわゆる全変動(ST)）は、$\sum_i(Y_i-\hat{Y}_i)^2$（標本観測値とモデルによる予測値との差の二乗和、いわゆる回帰からの残差の変動(SE)）と、$\sum_i(\hat{Y}_i-\bar{Y})^2$（線形回帰による変動(SR)）の和となっていることが分かる。モデルの説明力を高めるには ST に占める SR（SE）を極力大きく（小さく）することである。つまり、下式の決定係数（寄与率ともいう）R^2を定義し評価することになる。

$$R^2=\frac{SR}{ST}=\frac{\sum_i(\hat{Y}_i-\bar{Y})^2}{\sum_i(Y_i-\bar{Y})^2}=1-\frac{\sum_i(Y_i-\hat{Y}_i)^2}{\sum_i(Y_i-\bar{Y})^2}$$

この決定係数は、下式の通り被説明変数とその予測値との相関係数の２乗をとったものでもある。

$$R^2=\frac{S_{Y\hat{Y}}^2}{S_{YY}S_{\hat{Y}\hat{Y}}}=\frac{(\sum(Y_i-\bar{Y})(\hat{Y}_i-\bar{Y}))^2}{\sum(Y_i-\bar{Y})^2\sum(\hat{Y}_i-\bar{Y})^2}=\frac{\left\{\sum\left\{(\hat{Y}_i-\bar{Y})+(Y_i-\hat{Y}_i)\right\}(\hat{Y}_i-\bar{Y})\right\}^2}{\sum(Y_i-\bar{Y})^2\sum(\hat{Y}_i-\bar{Y})^2}=*$$

$$\sum_i(Y_i-\hat{Y}_i)(\hat{Y}_i-\bar{Y})=\sum_i\hat{\varepsilon}_i(\hat{Y}_i-\bar{Y})=0$$

$$*=\frac{\left\{\sum(\hat{Y}_i-\bar{Y})^2\right\}^2}{\sum(Y_i-\bar{Y})^2\sum(\hat{Y}_i-\bar{Y})^2}=\frac{\sum(\hat{Y}_i-\bar{Y})^2}{\sum(Y_i-\bar{Y})^2}=\frac{SR}{ST}=\frac{ST-SE}{ST}=1-\frac{SE}{ST}=1-\frac{\sum\hat{\varepsilon}_i^2}{\sum(Y_i-\bar{Y})^2}$$

決定係数の平方根Rを単回帰モデルでの相関係数に準じて重回帰モデルなので重相関係数と呼ぶ。単回帰モデルの相関係数は下記の通り重相関係数で求めても同じである。

$$R=\frac{\sum(Y_i-\bar{Y})(\hat{Y}_i-\bar{Y})}{\sqrt{\sum(Y_i-\bar{Y})^2\sum(\hat{Y}_i-\bar{Y})^2}}=\frac{\beta_1\sum_i(Y_i-\bar{Y})(X_i-\bar{X}))}{\beta_1\sqrt{\sum_i(Y_i-\bar{Y})^2}\sqrt{\sum_i(X_i-\bar{X})^2}}=\frac{\sum_i(Y_i-\bar{Y})(X_i-\bar{X}))}{\sqrt{\sum(Y_i-\bar{Y})^2}\sqrt{\sum_i X_i-\bar{X})^2}}=r$$

なお、決定係数は被説明変数の変動量に占める推計誤差変動量で定義しているが、正確には標本観測データに依らざるを得ないため、詳細は触れないが自由度を考慮した不偏分散を用いて自由度で修正した決定係数$\bar{R}^2$を評価することになる。これの平方根は自由度で修正した重相関係数$\bar{R}$となる。

$$\bar{R}^2=1-\frac{V_E}{V_T}=1-\frac{\sum\hat{\varepsilon}_i^2/(n-k-1)}{\sum(Y_i-\bar{Y})^2/(n-1)}$$

重相関係数が高いからと言ってモデル式の信頼性が高いとは必ずしも言えないし、どの程度高ければ良しとするかという問題もある。また異なったモデル間の推計精度の比較もできない。そこで、モデル式の信頼性を検定する方法がある。
先の誤差項に関する 4 個の仮定のもとにおいて、下式で表される分散比 F は自由度 $(k, n-k-1)$ の F 分布をすることが知られている。

$$F = \frac{V_R}{V_E} = \frac{SR/k}{SE/(n-k-1)}$$

したがってこれを用いて重回帰式の有意性を

帰無仮説 H_0 ：求めた重回帰式は Y の推定に何ら役に立たない

のもとで、有意水準 α %に対する F 分布表の $F(k, n-k-1; \alpha)$ に対応する値と、$F = V_R/V_E$ の値を比較して、もし $F = V_R/V_E > F(k, n-k-1; \alpha)$ ならば、危険率 α %で H_0 を棄却する（重回帰式を受け入れる）。

モデル式全体の検定のほかに、偏回帰係数が有効なのかどうかを検定することも重要である。つまり、

帰無仮説 H_0 ： $\beta_j = 0 \quad (j = 1, \cdots, k)$

を、それに対立する仮説（対立仮説）

対立仮説 H_1 ： $\beta_j \neq 0 \quad (j = 1, \cdots, k)$

に対して棄却できるかどうかを検定する。つまり、

$$t_j = \frac{|\hat{\beta}_j|}{\sigma\sqrt{V_j}} \quad (j = 1, \cdots, k)$$

の値と t 分布表の $t_\alpha(n-k-1; \alpha)$ の値を比較して、もし $t_j \geq t_\alpha(n-k-1; \alpha)$ ならば、危険率 α %で H_0 を棄却する。なお実際には、σ は既知でないことが殆どなので $\hat{s}^2 = \sum \hat{\varepsilon}_i^2/(n-k-1)$ を使って、$\sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$ の代わりに $\hat{s}^2(\mathbf{X}^t\mathbf{X})^{-1}$ を用いて、下式が自由度 $(n-k-1)$ の t 分布にしたがうことを利用する。

$$t_{j,n-k-1} = \frac{|\hat{\beta}_j|}{\hat{s}\sqrt{V_j}} \quad (j = 1, \cdots, k)$$

なお、V_j $(j = 1, \cdots, k)$ は $(\mathbf{X}^t\mathbf{X})^{-1}$ の j 行 j 列の対角要素である。

例題

人口が多く可住地が広い地域ほど車で移動する人数が多いのではないかと想像される。そこで都道府県別のデータからこの関係を回帰分析で試してみよう。なお線型重回帰モデル

に使用する変数（データベクトル）は旅客輸送人員を$\mathbf{Y}$、説明変数のうち定数項（値＝１）を$\mathbf{X}_1$、人口を$\mathbf{X}_2$、可住地面積を$\mathbf{X}_3$とし、説明変数のデータ行列を$\mathbf{X}=[\mathbf{X}_1 \quad \mathbf{X}_2 \quad \mathbf{X}_3]$とする。

表－１　都道府県別旅客輸送人員、人口及び可住地面積　（資料　平成20年　国土交通省　他）

		旅客輸送人員（バス+自家用車）単位：100万人	X_2：人口 単位：万人	X_3：可住地面積 単位：100Km^2	実測値Yの変動	回帰による予測値	回帰からの残差	回帰からの残差の変動	回帰による変動
		$\mathbf{Y}$	$\mathbf{X}_2$	$\mathbf{X}_3$	$(\mathbf{Y}-\overline{\mathbf{Y}})^2$	$\hat{\mathbf{Y}}$	$\mathbf{Y}-\hat{\mathbf{Y}}$	$(\mathbf{Y}-\hat{\mathbf{Y}})^2$	$(\hat{\mathbf{Y}}-\overline{\mathbf{Y}})^2$
01	北海道	2232	554	232.84503	1697199.14	2396.487948	−164.4879484	27056.28517	2152834.024
02	青森県	471	139	32.08738	209978.4378	578.3386052	−107.3386052	11521.57617	123127.608
03	岩手県	474	135	37.12797	207238.0335	583.7188026	−109.7188026	12038.21565	119380.781
04	宮城県	911	234	31.3298	332.4803078	841.154654	69.84534597	4878.372353	7757.978682
05	秋田県	377	111	31.58464	304962.4378	498.4931101	−121.4931101	14760.57581	185537.7509
06	山形県	409	119	28.52838	270643.459	510.7991754	−101.7991754	10363.07212	175087.738
07	福島県	859	205	42.17598	4932.820733	795.7808871	63.21911285	3996.65623	17809.74469
08	茨城県	1355	296	39.74592	181276.6505	1041.934639	313.0653613	98009.92042	12701.42438
09	栃木県	900	201	29.4768	854.629244	742.9122081	157.0877919	24676.57437	34715.826
10	群馬県	855	201	22.97043	5510.693074	721.5484169	133.4515831	17809.32504	43133.31912
11	埼玉県	2356	711	25.66772	2035661.097	2154.666331	201.3336695	40535.24646	1501684.292
12	千葉県	2115	612	34.86132	1406040.906	1908.379439	206.6205611	42692.05626	958725.7072
13	東京都	3044	1,284	13.95944	4472235.055	3716.421439	−672.4214395	452150.5923	7768413.586
14	神奈川県	2861	892	14.59264	3731719.714	2623.774224	237.2257764	56276.06897	2871466.425
15	新潟県	781	239	44.79548	21973.33137	899.3328199	−118.3328199	14002.65625	894.083119
16	富山県	408	110	18.52128	271684.9271	452.8066467	−44.80664673	2007.635591	226983.0635
17	石川県	454	117	13.8138	225847.3952	456.8982301	−2.898230126	8.399737865	223101.1197
18	福井県	317	81	10.6845	374830.5229	346.0870208	−29.08702083	846.0547809	340060.4489
19	山梨県	348	87	9.51045	337833.0122	358.9880205	−10.98802046	120.7365935	325180.5257
20	長野県	899	217	33.2269	914.0973291	799.9084414	99.09155858	9819.136981	16725.11111
21	岐阜県	904	210	21.45442	636.7569036	741.7045919	162.2954081	26339.79948	35167.29484
22	静岡県	1476	380	27.3078	298953.0122	1235.678048	240.3219521	57754.64065	93907.92843
23	愛知県	3270	740	29.59545	5479185.268	2248.550515	1021.449485	1043359.049	1740595.956
24	三重県	739	188	20.2195	36188.99095	676.2109852	62.78901484	3942.460384	64020.66757
25	滋賀県	523	140	12.89457	165026.0973	518.1113131	4.888686872	23.89925933	169021.8986
26	京都府	831	263	11.5325	9649.927116	857.137246	−26.13724599	683.1556281	5197.948074
27	大阪府	1999	881	13.1911	1144399.204	2588.452878	−589.4528785	347454.6959	2753007.145
28	兵庫県	1656	559	27.62284	528188.7569	1736.600281	−80.60028095	6496.405289	651840.2429
29	奈良県	465	140	8.4893	215513.2463	503.6465242	−38.64652423	1493.553835	181124.7357
30	和歌山県	296	101	10.96432	400985.3526	402.8591999	−106.8591999	11418.8886	277070.475
31	鳥取県	191	60	9.1182	544989.5016	282.2979907	−91.29799067	8335.323101	418526.2552
32	島根県	208	72	12.54396	520178.5441	327.0585704	−119.0585704	14174.94319	362615.2992
33	岡山県	710	195	22.12143	48063.56541	702.0046936	7.995306401	63.92492445	51633.17703
34	広島県	1003	287	22.55414	5441.416478	960.3510749	42.64892513	1818.930815	968.2697
35	山口県	497	146	17.54718	186826.2675	550.1442631	−53.14426309	2824.312699	143709.0609
36	徳島県	291	79	10.24309	407342.6931	339.0523041	−48.05230412	2309.023932	348314.4844
37	香川県	355	100	9.92933	329744.7356	396.6681211	−41.66812106	1736.232313	283626.4607
38	愛媛県	432	144	16.69332	247241.6931	541.7552589	−109.7552589	12046.21685	150139.8078
39	高知県	218	77	11.6522	505853.8633	338.0938055	−120.0938055	14422.52211	349446.7799
40	福岡県	1895	505	27.42327	932703.8846	1585.14085	309.8591497	96012.69265	430213.7405
41	佐賀県	318	86	13.3956	373607.0548	368.9523161	−50.95231608	2596.138514	313915.613
42	長崎県	437	144	16.29288	242294.3526	540.4404064	−103.4404064	10699.91768	151160.4915
43	熊本県	690	182	27.47626	57232.92712	683.282681	6.717319016	45.12237476	60492.07226
44	大分県	466	120	17.6886	214585.7782	477.9992183	−11.99921826	143.981239	203612.8667
45	宮崎県	421	114	18.33195	258301.842	463.355655	−42.35565497	1794.001508	217042.672
46	鹿児島県	582	172	32.43717	120571.4803	671.6452322	−89.64523217	8036.26765	66351.99524
47	沖縄県	375	138	11.63036	307175.3739	508.3749168	−133.3749168	17788.86842	177122.4038
	平均	929.234043							
				計（変動量）	29342550.4			2537384.13	26805166.3
				自由度	46			44	2
				不偏分散	637881.531			57667.821	13402583.1

$$\mathbf{X}^t\mathbf{X} = \begin{bmatrix} 47 & 12768 & 1227.857 \\ 12768 & 6669018 & 405868.8 \\ 1227.857 & 405868.8 & 80176.03 \end{bmatrix}$$

$$(\mathbf{X}^t\mathbf{X})^{-1} = \begin{bmatrix} 0.052933 & -7.5E-05 & -0.00043 \\ -7.5E-05 & 3.23E-07 & -4.86248E-07 \\ -0.00043 & -4.86248E-07 & 2.15216E-05 \end{bmatrix} \mathbf{X}^t\mathbf{Y} = \begin{bmatrix} 43674 \\ 21039741 \\ 1500837 \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y} = \begin{bmatrix} 84.79805 \\ 2.792669 \\ 3.283519 \end{bmatrix}$$

偏回帰係数が求められたので、下式で予測値が計算できる。計算結果は表参照。

$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$

ここで重相関係数と相関係数は

$$R^2 = \frac{SR}{ST} = \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2} = \frac{26805166.3}{29342550.4} = 0.9135, \quad R = 0.95578$$

となって、相関が高い。

自由度で修正した重相関係数を計算しよう。実測値 n=47 サンプルの自由度は n-1=46、回帰からの残差の自由度は今回使用した変数の数 n-k-1=44 である。よって

$$\bar{R}^2 = 1 - \frac{\sum (Y_i - \hat{Y})^2/(n-k-1)}{\sum (Y_i - \bar{Y})^2/(n-1)} = 1 - \frac{57667.821}{637881.531} = 0.9096$$

また片側 5%確率で F 検定を行うと、次のようになり仮説は棄却されモデル式は採択となる。

$$F = \frac{V_R}{V_E} = \frac{\sum_i (\hat{Y}_i - \bar{Y})^2 \Big/ k}{\sum_i (Y_i - \hat{Y}_i)^2 \Big/ (n-k-1)} = \frac{13402583.1}{57667.821} = 232.41 > F(k=2, n-k-1=44; 0.05) = 3.21$$

次に、偏相関係数 t 検定を行う。

$$t_{1,44} = \frac{|\hat{\beta}_1|}{\hat{s}\sqrt{V_1}} = \frac{|84.79805|}{\sqrt{57667.821 \times 0.052933}} = 1.535 > t_{0.14}(n-k-1=44; \alpha=0.14) = 1.5029$$

$$t_{2,44} = \frac{|\hat{\beta}_2|}{\hat{s}\sqrt{V_2}} = \frac{|2.792669|}{\sqrt{57667.821 \times (3.23E-07)}} = 20.448 > t_{0.05}(n-k-1=44; \alpha=0.05) = 2.0154$$

$$t_{3,44} = \frac{|\hat{\beta}_3|}{\hat{s}\sqrt{V_3}} = \frac{|3.283519|}{\sqrt{57667.821 \times (2.15216E-05)}} = 2.947378 > t_{0.05}(n-k-1=44; \alpha=0.05) = 2.01$$

$\hat{\beta}_1$ は 5%確率の t 検定で棄却されず、14%確率で漸く棄却される結果となった。$\hat{\beta}_2$、$\hat{\beta}_3$ ともに 5%確率の t 検定で棄却される結果となった。このことから $\hat{\beta}_2$、$\hat{\beta}_3$ は有意水準 5%で採択

可能であり、$\hat{\beta}_1$ は採択に当たって信頼が高いとは言い難いものの採択を断念するほどの問題はないであろう。
以上のことから、旅客交通量は人口と可住地面積による線形回帰モデルで表現できる可能性がある、といって良いであろう。