2008 年度前期　応用統計学　第４回

回帰分析 **(3)** － 回帰係数の統計的推測

今回は，回帰の考え方と，標本からの統計的推測の考え方を結びつけた，回帰係数に関する統計的推測を行います．回帰は，散布図上のデータから説明変数と被説明変数の関係を方程式で表す方法です．しかし，実際には散布図上のデータが調べるべきデータの全てであるとは限りません．そこで，散布図上のデータがある確率分布にしたがう母集団から無作為抽出された標本であるとして，回帰係数の統計的推測，すなわち区間推定や検定を行います．

母回帰係数と標本回帰係数

第２回の講義で，都市の緯度と気温の関係を例にとって，散布図上のデータから最小二乗法によって回帰方程式を導くことを説明しました．この時は，散布図上にあるデータは調査対象のすべてのデータであると考え，「それ以外のデータ」のことは考えていませんでした．

しかし，実際には調査対象のすべてのデータを調べる全数調査が行なえないこともしばしばあります．そこで，今回は同じ緯度であっても世界にはいろいろな気温のいろいろな都市があり，今得られている都市の気温のデータは，その中から偶然選ばれたひとつの標本であると考えます．すなわち，図１のように，各々の都市の緯度 x_i に対して気温 y がしたがう確率分布が背景にあると考え，今データとして得られている気温 y_i はその確率分布にしたがう母集団から無作為抽出された「標本」だと考えるのです．

今回，y_i は，ある確率分布にしたがう母集団から得られた標本と考えていますから，y_i は「緯度が同じ x_i である都市のうちどれが選ばれるか」という偶然に左右されます．したがって，偶然選ばれて今手元にある y_i を用いて最小２乗法で求められた回帰直線も，やはり偶然に左右されることになります．

そこで，仮に，ある緯度のすべての都市の気温がわかり，偶然によらない「本当の」回帰直線が求められたとして，その回帰直線を $y = a + bx$ で表します．この方程式の a, b を母回帰係数といいます．これに対して，標本 y_i から最小２乗法で求められた回帰係数は，標本から母回帰係数を推定した量と考えられ，これを標本回帰係数といい，今回は $\hat{a}, \hat{b}$ で表します．母回帰係数は，１変量の統計的推測の場合でいえば，例えば母平均に相当し，母集団の中身がすべてわからなければ求められない（つまり，現実には求められない，神様だけが知っている）量です．これに対して，標本回帰係数は，１変量の統計的推測の場合でいえば標本平均に相当します．標本平均が確率変数であるのと同様，標本回帰係数も偶然選ばれた標本に左右される確率変数です．

１変量の統計的推測では，標本平均がしたがう確率分布をある条件下で求めて，母平均についての区間推定や検定を行いました．同様に，標本回帰係数がしたがう確率分布をある条件下で求めて，母回帰係数についての区間推定や検定を行うのが，今回説明する「回帰係数の統計的推測」です．

回帰係数の統計的推測

緯度が x_i のとき，「母回帰係数による回帰方程式によって x_i から導かれる気温 $a + bx_i$」と「標本として得られている気温 y_i の差」，すなわち $y_i - (a + bx_i)$ を e_i で表し，誤差項とよびます．y_i は偶然に左右される確率変数ですから，e_i も確率変数です．誤差項は，第２回で説明した残差 d_i と似ていますが，異なるものであることに注意してください．残差は，現に得られている標本 y_i と標本回帰係数から求めら

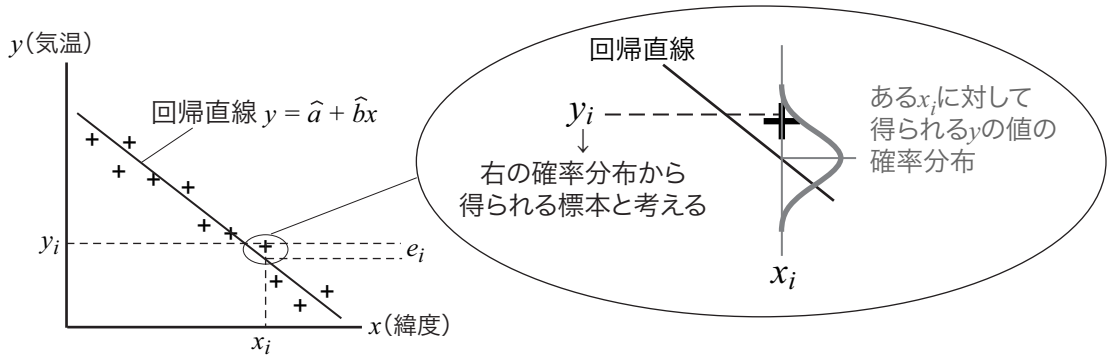

図 1: 回帰直線と確率分布

れる値で，今回の記号を用いれば $y_i - (\hat{a} + \hat{b}x_i)$ となります．

母回帰係数 a, b を推定するため，誤差項について次の仮定をします．

1. 各誤差項の期待値は 0（$E(e_i) = 0$）．つまり，標本 y_i は，母回帰係数から導かれる気温 $a + bx_i$ のまわりに偏りなく分布している，ということです．
2. 各誤差項の分散はどれも同じで σ^2（$V(e_i) = \sigma^2$）．
3. 各誤差項は無相関（$Cov(e_i, e_j) = 0 (i \neq j)$）．

次に，標本回帰係数 $\hat{a}, \hat{b}$ がしたがう確率分布を求めます．回帰係数で重要なのは回帰直線の傾きを決める b のほうなので，今回はこちらの確率分布を考えます．

まず，$\hat{b}$ の期待値 $E(\hat{b})$ を見てみましょう．第２回で示したことから，

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1}$$

です．ここで，

$$\bar{y} = \frac{\sum y_i}{n} = \frac{\sum(a + bx_i + e_i)}{n} = a + b\bar{x} \tag{2}$$

で，これらを (1) 式に代入すると，

$$\begin{aligned}
\hat{b} &= \frac{\sum(x_i - \bar{x})\left[a + bx_i + e_i - (a + b\bar{x} + \frac{\sum e_i}{n})\right]}{\sum(x_i - \bar{x})^2} \\
&= \frac{\sum(x_i - \bar{x})\left[b(x_i - \bar{x}) + (e_i - \frac{\sum e_i}{n})\right]}{\sum(x_i - \bar{x})^2} \\
&= b + \frac{\sum(x_i - \bar{x})e_i}{\sum(x_i - \bar{x})^2} - \left[\frac{\sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}\right]\frac{\sum e_i}{n}
\end{aligned} \tag{3}$$

が得られます．(3) 式を使って $\hat{b}$ の期待値 $E(\hat{b})$ を求めると，誤差項についての仮定 1. より $E(e_i) = 0$ ですから，右辺の第 2, 3 項はいずれも 0 となり，$E(\hat{b}) = E(b) = b$ となります．

さらに，$\hat{b}$ の分散 $V(\hat{b})$ を求めてみると，

$$\begin{aligned} V(\hat{b}) &= E\left[(\hat{b}-E(\hat{b}))^2\right] = E\left[(\hat{b}-b)^2\right] \\ &= E\left[\left(\frac{\sum(x_i-\bar{x})e_i}{\sum(x_i-\bar{x})^2}\right)^2\right] \quad \left(\frac{\sum e_i}{n}\text{を含む項は期待値をとると 0}\right) \\ &= E\left[\frac{\sum((x_i-\bar{x})^2e_i^2}{\left(\sum(x_i-\bar{x})^2\right)^2}\right] + E\left[\frac{\sum_i\sum_j(x_i-\bar{x})(x_j-\bar{x})e_ie_j}{\left(\sum(x_i-\bar{x})^2\right)^2}\right] \end{aligned} \tag{4}$$

となります（第 1 項は同番号の項の積，第 2 項は異番号の項の積．第 2 項では $i \neq j$ とします）．ここで仮定 3. より

$$Cov(e_i,e_j) = E\left[(e_i-E(e_i))\left(e_j-E(e_j)\right)\right] = E(e_ie_j) = 0 \tag{5}$$

です．また仮定 2. より $V(e_i)=\sigma^2$ で，$E(e_i)=0$ ですから

$$V(e_i) = E(e_i^2) - \{E(e_i)\}^2 = E(e_i^2) = \sigma^2 \tag{6}$$

となります．以上から，

$$V(\hat{b}) = \frac{\sigma^2}{\sum(x_i-\bar{x})^2} \tag{7}$$

となります（x_i は確率変数ではないので，期待値の計算においては定数であることに注意してください）．

さて，ここで各誤差項 e_i が独立で，いずれも正規分布 $N(0,\sigma^2)$ にしたがうと仮定しましょう．(3) 式で示されたように，$\hat{b}$ は e_i の 1 次関数になっていますから，$\hat{b}$ も正規分布にしたがいます．その平均と分散はここまでに求めた通りですから，$\hat{b}$ は

$$N\left(b, \frac{\sigma^2}{\sum(x_i-\bar{x})^2}\right) \tag{8}$$

にしたがいます．よって，

$$\frac{\hat{b}-b}{\sqrt{\frac{\sigma^2}{\sum(x_i-\bar{x})^2}}} \tag{9}$$

が標準正規分布 $N(0,1)$ にしたがうことがわかります．しかし，(9) 式では b と σ^2 の 2 つの未知数が入っているので，これで b の推定・検定を行うことはできません．そこで，「t 分布にもとづく推定・検定」[1]と同様に，σ^2 を標本から計算できる不偏分散で置き換えることを考えます．ここで，残差 d_i，すなわち

$$d_i = y_i - (\hat{a}+\hat{b}x_i) \tag{10}$$

を使って，

$$s^2 = \frac{\sum d_i^2}{n-2} \tag{11}$$

とおくと，この s^2 が自由度 $n-2$ の不偏分散[2]となることが示せます（証明は高度なので省略します）．1 変量の場合と同様に，(9) 式のように標準正規分布にしたがう確率変数の σ^2 を不偏分散で置き換えた確率変数を t 統計量といい，また t 統計量は不偏分散の自由度と同じ自由度の t 分布にしたがいます．この場合も，この s^2 で σ^2 を置き換えた

$$t_b = \frac{\hat{b}-b}{\sqrt{\frac{s^2}{\sum(x_i-\bar{x})^2}}} \tag{12}$$

[1] 2006 年度後期「情報統計学」第 15 回の講義録を参照してください．

[2]「不偏分散の自由度」とは，「平均からのへだたりの 2 乗の合計」を割る数のことです．1 変量データの分布の場合の不偏分散の自由度は $n-1$ です．くわしくは，2006 年度後期「情報統計学」第 8 回の講義録を参照してください．

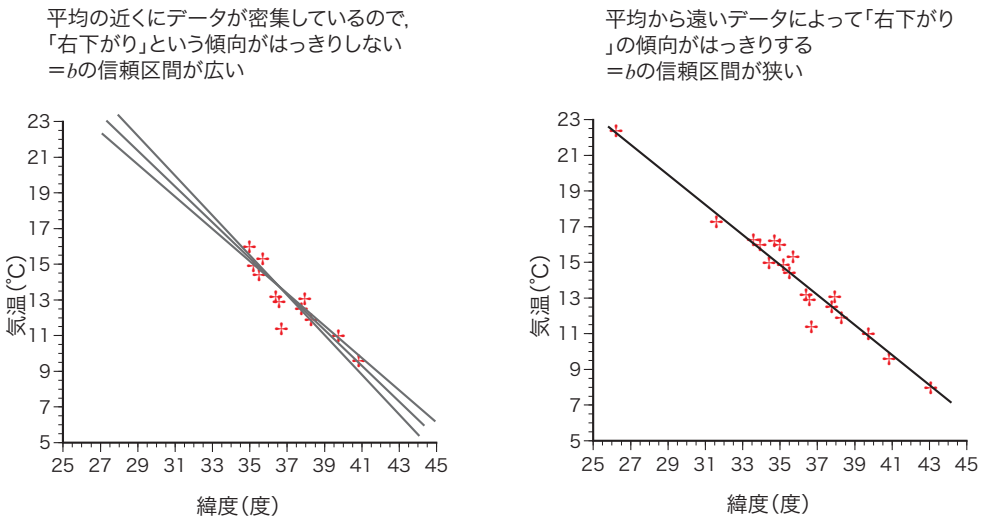

図 2: 平均に近いデータばかりの場合（左）と，平均から遠いデータがある場合（右）

は，自由度 $n-2$ の t 分布 $t(n-2)$ にしたがいます．なお，

$$s.e.(\hat{b}) = \sqrt{\frac{s^2}{\sum(x_i - \bar{x})^2}} \tag{13}$$

を $\hat{b}$ の標準誤差といいます．

これを用いると，回帰係数の信頼区間を求めたり，回帰係数についての検定を行うことができます．(12) 式の t_b は自由度 $n-2$ の t 分布 $t(n-2)$ にしたがいますから，$t_{0.025}(n-2)$ を「自由度 $n-2$ の t 分布において，t 統計量が $t_{0.025}(n-2)$ 以上である確率が 0.025 になるような t の値」とすると，

$$P(-t_{0.025}(n-2) \leq t_b \leq t_{0.025}(n-2)) = 0.95 \tag{14}$$

となります．これに (12) 式を代入すると，

$$P\left(\hat{b} - t_{0.025}(n-2)\sqrt{\frac{s^2}{\sum(x_i - \bar{x})^2}} \leq b \leq \hat{b} + t_{0.025}(n-2)\sqrt{\frac{s^2}{\sum(x_i - \bar{x})^2}}\right) = 0.95 \tag{15}$$

のように 95% 信頼区間が得られます．b についての検定も，同様に行えます．

ところで，(15) 式を見ると，不偏分散 s^2 が同じとき，$\sum(x_i - \bar{x})^2$ が大きいほど標準誤差は小さくなります．このことは，b の信頼区間が狭くなること，つまり推測が確実になることを意味しています．つまり，散布図上で，同じように直線の周りにデータが分布していても，平均から遠いデータが多いほど直線の傾きの値の不確実さが小さくなる，言いかえれば信頼度があがることを示しています．

図 2 は，第 2 回講義の演習問題の解答で示したもので，「都市の緯度と気温」のデータのうち，左は長野－鹿児島のデータだけから描いた散布図，右は札幌－那覇の全部のデータを用いたものです．この演習で求めたように，決定係数は左の場合が 0.712，右の場合が 0.949 です．この違いは，上で述べた「散布図上で平均から遠いデータが多いほど直線の傾きの信頼度があがる」ことに対応しています．