応用システム工学
第七回　回帰分析

平成25年07月19日

標本平均

大数の法則

- 標本数　→ 用いる母集団の数
- 標本サイズ→ 母集団からとりだす標本の数
- 大数の法則→ 標本サイズを大きくすると，その期待値が母集団の期待値から離れた値となる確率は小さくなる
  - 標本平均の期待値 $E\left[\overline{X}_n\right]$ と分散 $V\left[\overline{X}_n\right]$ で評価
  - 標本$X_1, X_2, \cdots, X_n$は母集団と同じ確率分布（平均μ，分散σ^2）となる確率変数
  - 標本$X_1, X_2, \cdots, X_n$は独立

母回帰係数の推定

- 仮定
  - 誤差の期待値 $E[e_i] = 0$
  - 誤差の分散 $V[e_i] = \sigma^2$
  - 誤差は無相関 $Cov[e_i, e_j] = 0 (i \neq j)$
  - 誤差は正規分布 $N(0, \sigma^2)$
- 標本回帰係数の期待値との関係を求める
  - 標本回帰係数 $\hat{a}_0, \hat{a}_1$

母回帰係数の推定

- 標本回帰係数 $\hat{a}_1$ の期待値 $E[\hat{a}_1]$

$$\hat{a}_1 = \frac{s_{xy}}{s_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$y_i = a_0 + a_1 x_i + e_i$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(a_0 + a_1 x_i + e_i)$$

$$= a_0 + a_1 \bar{x}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(a_0 + a_1 x_i + e_i - [a_0 + a_1 \bar{x}])}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(a_1[x_i - \bar{x}] + e_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = a_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})e_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

母回帰係数の推定

- 標本回帰係数 $\hat{a}_1$ の期待値$E[\hat{a}_1]$

$$E[\hat{a}_1] = E\left[a_1 + \frac{\sum_{i=1}^{n}(x_i - \overline{x})e_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right] = E[a_1] + E\left[\frac{\sum_{i=1}^{n}(x_i - \overline{x})e_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right]$$

  – 確率変数 e_i なので，第二項は0

$$E[\hat{a}_1] = E[a_1] = a_1$$

母回帰係数の推定

- 標本回帰係数 $\hat{a}_1$ の分散 $V[\hat{a}_1]$

$$V[\hat{a}_1] = E\left[\left(\hat{a}_1 - E[\hat{a}_1]\right)^2\right] = E\left[\left(\hat{a}_1 - a_1\right)^2\right]$$

$$= E\left[\left(a_1 + \frac{\sum_{i=1}^{n}(x_i - \overline{x})e_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} - a_1\right)^2\right]$$

$$= E\left[\left(\frac{\sum_{i=1}^{n}(x_i - \overline{x})e_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)^2\right]$$

母回帰係数の推定

- つづき

$$V[\hat{a}_1] = E\left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - \bar{x})(x_j - \bar{x})e_i e_j}{\left\{\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}^2}\right]$$

$$Cov[e_i, e_j] = E[(e_i - E[e_i])(e_j - E[e_j])] = E[e_i e_j] = 0$$

$$= E\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 e_i{}^2}{\left\{\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}^2}\right]$$

$$V[e_i] = E[e_i{}^2] - E[e_i]^2 = E[e_i{}^2] = \sigma^2$$

$$= \frac{\sigma^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{\left\{\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

母回帰係数の推定

- 標本回帰係数 $\hat{a}_1$ と母回帰係数 a_1 の関係

$$\hat{a}_1 = a_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})e_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- 標本回帰係数 $\hat{a}_1$ の期待値 $E[\hat{a}_1]$

$$E[\hat{a}_1] = E[a_1] = a_1$$

- 標本回帰係数 $\hat{a}_1$ の分散 $V[\hat{a}_1]$

$$V[\hat{a}_1] = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

母回帰係数の推定

- 標本回帰係数 $\hat{a}_1$ は誤差 e_i の一次関数として表される

$$\hat{a}_1 = a_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})e_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

→標本の誤差が正規分布なので標本回帰係数も正規分布する

$$\hat{a}_1 \sim N\left(a_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

– 標準正規分布化

$$\frac{\hat{a}_1 - a_1}{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \sim N(0,1)$$

- 期待値を引いて分散で割る

→ただし，σ^2，a_1は未知数

確率変数の性質

- 確率変数Xの関数g(X)の期待値E[g(X)]
  - 確率密度関数:確率変数Xが値xをとる確率f(x)

$$E[g(X)] = \sum_x g(x)f(x)$$

    - 確率変数の期待値E[X]→g(x)=x
(原点の周りの1次のモーメント(積率))
    - 確率変数の分散V[X]→g(x)=(x-μ)^2
(期待値の周りの2次のモーメント)
  - 定数cに対する性質
    - 期待値 $E[cX] = \sum_x cxf(x) = c\sum_x xf(x) = cE[X]$
    - 分散$V[cX] = \sum_x (cx - c\mu)^2 f(x) = c^2 \sum_x (x-\mu)^2 f(x) = c^2 V[X]$

同時確率分布

- 同時確率分布
  - 確率変数Xの値がxとなり，同時に確率変数Yの値がyとなる確率分布
  - X,Yの同時確率密度関数　$f_{XY}(x, y)$
- 周辺確率分布
  - 同時確率分布から，ある確率変数の分布を抽出したもの
  - 確率変数Xの周辺確率密度関数　$f_X(x)$

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy$$

複数の確率変数の性質

- 2つの確率変数の和に対する期待値

$$
\begin{aligned}
E[X+Y] &= \int_x\int_y (x+y) f_{XY}(x,y)dxdy \\
&= \int_x\int_y x f_{XY}(x,y)dxdy + \int_x\int_y y f_{XY}(x,y)dxdy \\
&= \int_x x\int_y f_{XY}(x,y)dydx + \int_y y\int_x f_{XY}(x,y)dxdy \\
&= \int_x x f_X(x)dx + \int_y y f_Y(y)dy \\
&= E[X]+E[Y]
\end{aligned}
$$

複数の確率変数の性質

- 確率変数X,Yが独立である場合における，
  同時確率密度関数$f_{XY}(x,y)$と，
  周辺確率密度関数 $f_X(x), f_Y(y)$ の関係

$$f_{XY}(x,y) = f_X(x) f_Y(y)$$

- 独立な確率変数の積の期待値

$$E[XY] = \int_x\int_y xyf_{XY}(x,y)dxdy = \int_x\int_y xf_X(x)yf_Y(y)dxdy$$

$$= \int_x xf_X(x)dx\int_y yf_Y(y)dy = E[X]E[Y]$$

  – この時共分散は

$$Cov[X,Y] = E[(X-E[X])(Y-E[Y])] = E[XY - E[X]Y - XE[Y] + E[X]E[Y]]$$

$$= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y]$$

$$= E[XY] - E[X]E[Y] = 0$$

複数の確率変数の性質

- 分散と期待値の関係

$$V[X] = E\left[(X - E[X])^2\right] = E\left[X^2 - 2XE[X] + E[X]^2\right]$$
$$= E\left[X^2\right] - 2E[X]E[X] + E[X]^2 = E\left[X^2\right] - E[X]^2$$

- 2つの確率変数の和に対する分散

$$V[X+Y] = E\left[(X+Y)^2\right] - E[X+Y]^2 = E\left[X^2 + 2XY + Y^2\right] - (E[X] + E[Y])^2$$
$$= E\left[X^2\right] + 2E[XY] + E\left[Y^2\right] - \left(E[X]^2 + 2E[X]E[Y] + E[Y]^2\right)$$
$$= \left(E\left[X^2\right] - E[X]^2\right) + \left(E\left[Y^2\right] - E[Y]^2\right) + 2(E[XY] - E[X]E[Y])$$

  – 2つの確率変数が独立の場合

$$V[X+Y] = V[X] + V[Y] + 2(E[X]E[Y] - E[X]E[Y])$$
$$= V[X] + V[Y]$$

標本平均

- 標本平均の期待値

$$E\left[\overline{X}_n\right]=E\left[\frac{X_1+X_2+\cdots+X_n}{n}\right]=\frac{1}{n}\left\{E\left[X_1\right]+E\left[X_2\right]+\cdots+E\left[X_n\right]\right\}$$

$$=\frac{1}{n}\left\{\mu+\mu+\cdots+\mu\right\}=\mu$$

- 標本平均の分散

$$V\left[\overline{X}_n\right]=V\left[\frac{X_1+X_2+\cdots+X_n}{n}\right]=\frac{1}{n^2}V\left[X_1+X_2+\cdots+X_n\right]$$

$$=\frac{1}{n^2}\left\{V\left[X_1\right]+V\left[X_2\right]+\cdots+V\left[X_n\right]\right\}=\frac{1}{n^2}\left\{\sigma^2+\sigma^2+\cdots+\sigma^2\right\}$$

$$=\frac{1}{n^2}n\sigma^2=\frac{1}{n}\sigma^2$$

→標本の数nが小さいと分散が大きい
標本の数を大きくすると期待値に近づく