応用システム工学
第七回　回帰分析

平成26年05月30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

標本と母数の関係
単回帰

- 母回帰係数 a_0, a_1
  - 予測値 $a_0 + a_1 x_i$
  - 標本 y_i, x_i
  - 誤差 $e_i = y_i - \left(a_0 + a_1 x_i\right)$
    (残差ではない)
- 標本回帰係数 $\hat{a}_0, \hat{a}_1$
  - 残差 $y_i - \left(\hat{a}_0 + \hat{a}_1 x_i\right)$
- 母集団を求めることは難しいので，標本から母集団を推定する

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　$V[\hat{a}_1]$

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

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

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

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 $\hat{a}_1$ 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 $V[\hat{a}_1]$

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

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 $\dfrac{\hat{a}_1 - a_1}{\dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \sim N(0,1)$
    - 期待値を引いて分散で割る

→ただし，σ^2，a_1は未知数

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

不変分散

- 一致推定量→標本サイズが十分大きいとき，標本平均が母平均に確率収束する性質をもつ推定量→標本平均は母平均に確率収束
- 不変推定量→期待値が母数に等しい推定量
  - 標本平均 $\overline{X}$ は不偏推定量
    - 母平均 μ に対して $E\left[\overline{X}\right]=\mu$

標本平均は
母平均と等しくなる

  - 標本の分散は不偏推定量となるか？
    - 標本の分散

$$S^2=\frac{\left(X_1-\overline{X}\right)^2+\left(X_2-\overline{X}\right)^2+\cdots+\left(X_n-\overline{X}\right)^2}{n}$$

不変分散

- 標本の分散

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[\left(X_i - \mu\right) - \left(\overline{X} - \mu\right)\right]^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \overline{Y}\right]^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}Y_i^2 - 2\frac{\overline{Y}}{n}\sum_{i=1}^{n}Y_i + \frac{\overline{Y}^2}{n}\sum_{i=1}^{n}1$$

$$= \frac{1}{n}\sum_{i=1}^{n}Y_i^2 - 2\overline{Y}\overline{Y} + \frac{\overline{Y}^2 n}{n} = \frac{1}{n}\sum_{i=1}^{n}Y_i^2 - \overline{Y}^2$$

変数変換

$$Y_i = X_i - \mu$$

$$\overline{Y} = \overline{X} - \mu$$

不変分散

- 標本の分散の期待値

$$E\left[S^2\right]=E\left[\frac{1}{n}\sum_{i=1}^{n}Y_i^2-\overline{Y}^2\right]=E\left[\frac{1}{n}\sum_{i=1}^{n}Y_i^2\right]-E\left[\overline{Y}^2\right]$$

$$=\frac{1}{n}\sum_{i=1}^{n}E\left[Y_i^2\right]-E\left[\overline{Y}^2\right]=\frac{1}{n}\sum_{i=1}^{n}E\left[\left(X_i-\mu\right)^2\right]-E\left[\left(\overline{X}-\mu\right)^2\right]$$

  – 標本サイズが大きいときの標本と母平均の差の二乗平均

$$E\left[\left(X_i-\mu\right)^2\right]\rightarrow\sigma^2$$

  – 標本平均の分散

$$E\left[\left(\overline{X}-\mu\right)^2\right]=\frac{\sigma^2}{n}$$

不変分散

- 標本分散の期待値

$$E\left[S^2\right]=\frac{1}{n}\sum_{i=1}^{n}\sigma^2-\frac{\sigma^2}{n}$$

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

  – 標本分散 S^2 の期待値$E\left[S^2\right]$は母分散 σ^2とは等しくない
    →標本分散S^2は母分散の不偏推定量ではない
    $\frac{n}{n-1}S^2$ が母分散の不偏推定量に相当

- 不偏分散（自由度n-1）

$$s^2=\frac{\left(X_1-\overline{X}\right)^2+\left(X_2-\overline{X}\right)^2+\cdots+\left(X_n-\overline{X}\right)^2}{n-1}$$