

以下はある造船所で建造されたタンカーの主要目とその工事に伴う溶接長と塗料量の実績である。

| 船の番号 | L | B | D | Cb | 延べ溶接長 | 塗料合計 |
|------|-----|------|------|------|-------|------|
| No.1 | 168 | 32.2 | 17.0 | 0.80 | 11.3 | 320 |
| No.2 | 192 | 32.2 | 15.2 | 0.79 | 12.4 | 350 |
| No.3 | 116 | 20.0 | 13.7 | 0.78 | 3.32 | 89.7 |
| No.4 | 175 | 23.1 | 15.2 | 0.82 | 11.7 | 332 |
| No.5 | 137 | 23.1 | 14.5 | 0.78 | 8.72 | 247 |
| No.6 | 172 | 32.2 | 18.1 | 0.80 | 12.3 | 349 |
| No.7 | 97 | 15.5 | 10.5 | 0.70 | 2.1 | 51.2 |

【問 1】新しい工事予定の船の要目 L, B, D, Cb が与えられたとき、上記のデータを利用して、この工事予定の船における延べ溶接長と塗料合計を多重回帰を用いて推定したい。

上記の表データの数字を用いて式を作り、計算手順を説明せよ。(数値計算はしなくて良い)

【問 2】問 1 の推定をより精度良く行うため、以下のような項目 1,2,3 のような工夫を検討した。

このとき、明らかに推定計算上不都合が生じると考えられるものを全て挙げ、その理由を説明せよ。ただし単なる計算量の増加は、計算上の不都合とは考えないものとする。

1. 新しく番号 No.8 についてのデータをテーブルに追加し、問 1-1 と同様の推定方法で計算をやり直す。このとき、この No.8 は No.6 と同型船であるため、L, B, D, Cb の値は No.6 と完全に同一であるが、延べ溶接長と塗料合計は若干異なっている。
2. 各船の特徴量として L, B, D, Cb に加えて、船殻の体積を反映する新しい特徴量 $x_5 = L \cdot B \cdot D \cdot C_b$ を導入し、問 1-1 と同様の推定方法で計算をやり直す。
3. 各船の特徴量として L, B, D, Cb に加えて、船殻の長さを反映する新しい特徴量 $x_5 = L + B + D$ を導入し、問 1-1 と同様の推定方法で計算をやり直す。

海事統計学 第2回 相関/回帰 演習問題解答

【問1】各船の特徴量 L, B, D, C_b を以下の変数 (説明変数) x_1, x_2, x_3, x_4 で表し、延べ溶接長 y を以下の多重回帰によって説明する：

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + e$$

ただし b_0, b_1, b_2, b_3, b_4 は回帰係数、 e は確率変動 (誤差) である。

ここで、表のデータを用いて、回帰係数 b_0, b_1, \dots, b_4 を求める。ここで、

$$\mathbf{y} = \begin{bmatrix} 11.3 \\ 12.4 \\ 3.32 \\ 11.7 \\ 8.72 \\ 12.3 \\ 2.1 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & 168 & 32.2 & 17.0 & 0.80 \\ 1 & 192 & 32.2 & 15.2 & 0.79 \\ 1 & 116 & 20.0 & 13.7 & 0.78 \\ 1 & 175 & 23.1 & 15.2 & 0.82 \\ 1 & 137 & 23.1 & 14.5 & 0.78 \\ 1 & 172 & 32.2 & 18.1 & 0.80 \\ 1 & 97 & 15.5 & 10.5 & 0.70 \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} \quad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \end{bmatrix}$$

と表すと、 $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$ とした場合の誤差ベクトル \mathbf{e} の平方和を最小にする回帰係数 $\hat{\mathbf{b}}$ は以下の式で与えられる：

$$\hat{\mathbf{b}} = (\mathbf{X}^{Trans} \mathbf{X})^{-1} \mathbf{X}^{Trans} \mathbf{y}$$

新しい工事予定の船の要目 L, B, D, C_b を $x_1 = L, x_2 = B, x_3 = D, x_4 = C_b$ へ代入し、上で求めた回帰係数 $\hat{\mathbf{b}}$ を使って $y = \hat{b}_0 + \hat{b}_1x_1 + \hat{b}_2x_2 + \hat{b}_3x_3 + \hat{b}_4x_4$ より新しい工事予定の船の延べ溶接長の推定値 y を得る。

また、塗料合計については上記の行列のうち \mathbf{y} だけを以下のように

$$\mathbf{y} = \begin{bmatrix} 320 \\ 350 \\ 89.7 \\ 332 \\ 247 \\ 349 \\ 51.2 \end{bmatrix}$$

として同様の計算により推定する。

【問2】説明変数 x_1, x_2, \dots, x_k について、任意の定数 a_0, a_1, \dots, a_k を用いて関係式 $a_0 + a_1x_1 + a_2x_2 + \dots + a_kx_k = 0$ に近い関係が成り立つとき、データは強い多重共線関係 (多重共線性) があるという。このとき回帰係数ベクトルを求める逆行列計算が不安定になり、無意味な解が出やすい。

ここで、問題文のうち、明らかに3番目の方法は上記の関係式を生み出してしまい、回帰係数ベクトルを求める逆行列計算ができないという不都合が生じる。

多重回帰は最小2乗法であり、データの重複は問題ないので、1番目の方法は問題ない。また2番目の方法は、新しい説明変数を既存の説明変数から生成しているが、非線形関数になっているので多重共線性は生じず、問題は無い。