データ解析 2004 年度期末試験問題

静岡大学工学部システム工学科

安藤 和敏

2005 年 2 月 16 日

注意事項

- A4 片面に手書きのカンニングペーパー, 及び, (関数) 電卓のみ持ち込み可.
- 印刷あるいはコピーしたカンニングペーパーやノートパソコンの持ち込みは不可.
- 携帯電話の電卓機能の使用ももちろん不可.
- 試験の時間は 10:20-11:40 である.
- 数値計算の結果は, 丸めない (四捨五入しない) で記入すること. ただし, 採点の対象は有効数字 4 桁くらいまでである.
- また, 計算途中の数値を丸めてしまうと, 最終的な計算結果が正確でなくなるため, 計算途中の数値は丸めないこと.
- 問題用紙は持ち帰ってよい.
- 解答及び採点の結果は, Web ページ (`http://coconut.sys.eng.shizuoka.ac.jp/data/`) で公開する.

問題 1. (配点 26)

2 変数 x, y に関する n 個のデータ (x_i, y_i) $(i = 1, \cdots, n)$ が与えられたとき, 単回帰モデルは, y_i と x_i の間に

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1, \cdots, n) \tag{1}$$

という関係を仮定する. ここで, ε_i は互いに独立に期待値 0 分散 σ^2 の正規分布にしたがう確率変数である.

以下では, 単回帰モデルを仮定する. 以下の事実を用いて, β_1 の信頼区間を求めてみよう.

(i) 正規分布にしたがう確率変数の線形結合は正規分布にしたがう確率変数である.

(ii) 正規分布にしたがう確率変数を標準化すると標準正規分布にしたがう確率変数となる.

(iii) 任意の n 個の確率変数 z_i $(i = 1, \cdots, n)$ と定数 a_i $(i = 0, 1, \cdots, n)$ に対して $E(a_0 + \sum_{i=1}^n a_i z_i) = \sum_{i=1}^n a_i E(z_i)$.

(iv) 任意の n 個の互いに独立な確率変数 z_i $(i = 1, \cdots, n)$ と定数 a_i $(i = 0, 1, \cdots, n)$ に対して $V(a_0 + \sum_{i=1}^n a_i z_i) = \sum_{i=1}^n a_i^2 V(z_i)$.

以下の文章中の空欄 $\boxed{\text{ア}}$, $\boxed{\text{イ}}$, $\boxed{\text{エ}}$, $\boxed{\text{キ}}$, $\boxed{\text{ク}}$, $\boxed{\text{コ}}$, $\boxed{\text{シ}}$ の中にはもっとも良くあてはまる命題を (i)〜(iv) から選び, これら以外の空欄には最も適切な数式または語句を記入せよ.

設問 (1) 式 (1) によって各 y_i $(i = 1, \cdots, n)$ も確率変数になる. $\boxed{\text{ア}}$ によって y_i のしたがう分布は正規分布である. 各 $i = 1, \cdots, n$ に対して, $\boxed{\text{イ}}$ より $E(y_i) = \boxed{\text{ウ}}$ であり, $\boxed{\text{エ}}$ より $V(y_i) = \boxed{\text{オ}}$ である. したがって,

$$y_i \sim N(\boxed{\text{ウ}}, \boxed{\text{オ}}) \quad (i = 1, \cdots, n). \tag{2}$$

また, $i \neq j$ に対して, $Cov(y_i, y_j) = 0$ であることも簡単に確かめられるので, y_i と y_j は $\boxed{\text{カ}}$ である.

設問 (2) β_1 の最小 2 乗推定量 $\hat{\beta}_1$ は, $\hat{\beta}_1 = S_{xy}/S_{xx}$ で与えられるが, これを y_i の線形結合として,

$$\hat{\beta}_1 = \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) y_i \tag{3}$$

と書き表わすことができる. $\boxed{\text{キ}}$ によって,

$$E(\hat{\beta}_1) = \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) E(y_i) \tag{4}$$

は, $\boxed{\quad ク \quad}$ より正規分布にしたがう確率変数になる. さらに, 式 (4) は式 (2) によって

$$E(\hat{\beta}_1) = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x})(\boxed{\quad ウ \quad}) \tag{5}$$

となり, 式 (5) の右辺を計算すると最終的には $E(\hat{\beta}_1) = \boxed{\quad ケ \quad}$ を得る.

y_i $(i = 1, \cdots, n)$ は $\boxed{\quad カ \quad}$ であるから, 式 (3) と $\boxed{\quad コ \quad}$ によって,

$$V(\hat{\beta}_1) = \frac{1}{S_{xx}^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 V(y_i) \tag{6}$$

さらに, 式 (6) は式 (2) によって

$$\begin{aligned} V(\hat{\beta}_1) &= \frac{1}{S_{xx}^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 \boxed{\quad オ \quad} \\ &= \boxed{\quad サ \quad} \end{aligned}$$

を得る.

まとめると,

$$\hat{\beta}_1 \sim N(\boxed{\quad ケ \quad} , \boxed{\quad サ \quad})$$

である.

設問 (3) $\boxed{\quad シ \quad}$ によって, $\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 / S_{xx}}} \sim N(0, 1)$ を得る. ここで, $\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 / S_{xx}}}$ において, σ^2 のかわりにその推定量 $V_e = S_e/(n-2)$ を代入した $\frac{\hat{\beta}_1 - \beta_1}{\sqrt{V_e / S_{xx}}}$ は, 自由度 $n-2$ の t 分布にしたがう確率変数となる.

したがって, $t(\phi, 0.05)$ を自由度 ϕ の t 分布の両側 5%点とすると,

$$-t(n-2, 0.05) \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{V_e / S_{xx}}} \leq t(n-2, 0.05)$$

となる確率は, 0.95 である.

上式を書き直せば, β_1 の信頼率 95%の信頼区間は, $\boxed{\quad ス \quad}$ であることがわかる.

問題 2. (配点 26)

表 1 に示すような 2 変数 x, y についての $n = 10$ 個のデータが得られているとする.

表 1: 2 変数 x と y に関するデータ

No	x	y
1	28	-62
2	39	-100
3	25	-63
4	42	-110
5	42	-107
6	26	-62
7	11	-19
8	30	-73
9	16	-31
10	24	-52

この表から S_{xx}, S_{yy}, S_{xy} を求めると $S_{xx} = 978.1, S_{yy} = 8456.9, S_{xy} = -2864.3$ となる.

以下の設問に答えよ. $t(\phi, P)$ の値には, 表 2 の数値を用いよ.

設問 (1) 残差平方和 S_e の最小値を S_{yy}, S_{xx}, S_{xy} を使った式で表現し, その値を与えよ.

設問 (2) β_1 の 95%信頼区間, 及び, 99%信頼区間を数値で与えよ.

設問 (3) x_0 に対する母回帰 $\beta_0 + \beta_1 x_0$ の 95%信頼区間は,

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t(\phi_e, 0.05)\sqrt{\left\{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right\} V_e}$$

で与えられる. $x_0 = 20$ に対する母回帰 $\beta_0 + \beta_1 x_0$ の 95%信頼区間を数値で求めよ.

表 2: $t(\phi, P)$ の表

	P					
ϕ	0.20	0.10	0.05	0.02	0.01	0.001
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073

問題 3. (配点 20)

表 3 に示すような 3 変数 x_1, x_2, y についてのデータが得られているとする．このデータに対して, 重回

表 3: 3 変数 x_1, x_2, y に関するデータ

No	x_1	x_2	y
1	x_{11}	x_{12}	y_1
2	x_{21}	x_{22}	y_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	x_{i1}	x_{i2}	y_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{n1}	x_{n2}	y_n

帰モデル

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \ (i = 1, \cdots, n)$$

を仮定する．ここで, $\varepsilon_i \ (i = 1, \cdots, n)$ は独立に同一の分布 $N(0, \sigma^2)$ に従う確率変数である．$\beta_0, \beta_1, \beta_2$ の最小 2 乗推定量を, それぞれ, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ とするとき, 以下の設問に答えよ．

設問 (1) S_e は, 残差

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} \tag{7}$$

の平方和:

$$S_e = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \{y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}\}^2$$

として定義される．S_e を最小にする $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ は, 方程式

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) = 0, \tag{8}$$

$$\sum_{i=1}^{n} x_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) = 0, \tag{9}$$

$$\sum_{i=1}^{n} x_{i2}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) = 0 \tag{10}$$

の解である．式 (7), 式 (8)〜式 (10) を用いて,

$$\sum_{i=1}^{n} e_i \{(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}) - \bar{y}\} = 0 \tag{11}$$

を証明せよ．

設問 (2) 式 (11) を用いて,

$$S_{yy} = S_e + \sum_{i=1}^{n} \{(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}) - \bar{y}\}^2 \tag{12}$$

を証明せよ．

問題 4. (配点 28)

表 4 に示すような 3 変数 x_1, x_2, y に関するデータが得られているとする.

表 4: x_1, x_2, y に関するデータ

No.	x_1	x_2	y
1	94	58	38
2	57	43	44
3	78	47	32
4	61	45	56
5	66	43	32
6	93	32	-42
7	71	36	2
8	53	53	95
9	88	30	-42
10	88	57	41

このデータに対して, 重回帰モデル

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \ (i = 1, \cdots, 10)$$

を仮定する. ここで, $\varepsilon_i \ (i = 1, \cdots, 10)$ は独立に同一の分布 $N(0, \sigma^2)$ に従う確率変数である.

表 4 のデータに関する $S_{11}, S_{22}, S_{yy}, S_{12}, S_{1y}, S_{2y}$ を計算すると, 次の表 5 のようになる.

表 5: 平方和と偏差積和

S_{11}	S_{22}	S_{yy}	S_{12}	S_{1y}	S_{2y}
2132.9	860.4	16248.4	-106.6	-3887.4	2969.6

以下の設問に答えよ.

設問 (1) $\beta_0, \beta_1, \beta_2$ の最小 2 乗推定量 $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ を数値で与えよ.

設問 (2) 残差平方和 S_e, 回帰による変動 S_R を数値で与えよ.

設問 (3) 寄与率 R^2 と自由度調整済み寄与率 R^{*2} を数値で与えよ.

学籍番号		氏名	

問題1の解答欄 (配点各2)

ア. ______ イ. ______ ウ. ______

エ. ______ オ. ______ カ. ______

キ. ______ ク. ______ ケ. ______

コ. ______ サ. ______ シ. ______

ス. ______

問題2(1) の解答欄 (配点6)

問題2(2) の解答欄 (配点12)

95%信頼区間: ______

99%信頼区間: ______

問題2(3) の解答欄 (配点8)

問題 3(1) の解答欄 (配点 10)

問題 3(2) の解答欄 (配点 10)

問題 4(1) の解答欄 (配点各 4)

$\hat{\beta}_0 =$ ________________ $\hat{\beta}_1 =$ ________________ $\hat{\beta}_2 =$ ________________

問題 4(2) の解答欄 (配点各 4)

$S_e =$ ________________ $S_R =$ ________________

問題 4(3) の解答欄 (配点各 4)

$R^2 =$ ________________ $R^{*2} =$ ________________