RSS Higher Certificate in Statistics, Specimen A

Module 4 : Linear Models

1. 以下の各問に答えよ.

(i) 単回帰分析におけるモデルと標準的な仮定を示せ.

(ii) (a) 単回帰モデルにおいて切片項が 0 であることが分かっていて，モデルは

$$y_i = \beta x_i + e_i \qquad (*)$$

で，e_i には通常の仮定が置かれるとする．このとき β の最小 2 乗推定量は $\sum x_i y_i \big/ \sum x_i^2$ で与えられることを示せ.

(b) 1 ヶ月の間ある大都市で 10 本の道路について調査が実施された．各道路では，無作為に選ばれた 1 時間の間観測が行われ，平均交通量 x_i（1 分あたりの車の数）とスピード違反の車の数 y_i が記録された ($i = 1, 2, \ldots, 10$). 以下のデータを適当なグラフに表わし，上述のモデル (*) が適切であるかどうかをコメントせよ．そのモデルをデータに当てはめて車の台数が 1 分あたり 20 台である道路におけるスピード違反の車の数の期待値を求めよ．（これ以上の計算すること無しに）モデルに切片項が含まれるべきであるかどうかについてコメントせよ.

Flow, x	5	5	5	10	10	15	25	25	30	50
Violations, y	2	1	1	4	2	5	8	2	5	10

2. ある大都市で人口密度に関する調査が行われた．この都市から無作為に 10 個の居住地区が選ばれ，市の中心部からの距離と人口密度（100 人／km^2）が記録された．以下の表はそれらのデータおよび各測定値の対数を示している.

distance, x (km)	*population density, y*	*log x*	*log y*
0.4	149	–0.916	5.004
1.0	141	0.000	4.949
3.1	102	1.131	4.625
4.5	46	1.504	3.829
4.7	72	1.548	4.277
6.5	40	1.872	3.689
7.3	23	1.988	3.135
8.2	15	2.104	2.708
9.7	7	2.272	1.946
11.7	5	2.460	1.609

(i) 以下の組み合わせのデータのプロットにより，直線で最もよく表わされる関係は次のいずれであるかを示せ.

(a) (x, y) (b) $(\log x, \log y)$ (c) $(x, \log y)$

(ii) 以下に示す回帰分析の結果の出力より，どの回帰直線が最もよいと考えるかを述べよ.

解答には，出力における統計量のいずれを参照したか，および (i) で作成したグラフとの関係について示せ．log x に対し log y を回帰させるのがよいと考えるか．もしそうでないならばそれは何故か．

(iii) 上問 (ii) で最もよいとしたモデルにつき，y を x の関数として表せ．

(iv) 選択したモデルを用いて，市の中心からの距離が 5km の地区での人口密度を推定せよ．

(v) モデルを用いて人口密度を予測する際に注意すべき点があればそれを示せ．

```
Regression Analysis: y versus x

The regression equation is   y = 140 - 14.0x

Predictor      Coef  SE Coef      T      P
Constant     139.70    11.12  12.56  0.000
x           -13.958    1.663  -8.39  0.000

S = 18.2834   R-Sq = 89.8%   R-Sq(adj) = 88.5%

Observation 10 has an unusually large positive residual



Regression Analysis: y versus logx

The regression equation is y = 127 - 48.0logx

Predictor      Coef  SE Coef      T      P
Constant    126.990    9.147  13.88  0.000
logx        -47.980    5.293  -9.07  0.000

S = 17.0492   R-Sq = 91.1%   R-Sq(adj) = 90.0%

Observation 1 has an unusually large negative residual



Regression Analysis: logy versus x

The regression equation is logy = 5.41 - 0.322x

Predictor       Coef  SE Coef       T      P
Constant      5.4133   0.1621   33.40  0.000
x          -0.32157  0.02425  -13.26  0.000

S = 0.266544   R-Sq = 95.6%   R-Sq(adj) = 95.1%
```

3. 3 つのタイプの時計の文字盤 (watch dial) が 21 人の被験者によってテストされた．各被験者には無作為に 1 つの文字盤が割り当てられ，標準的な実験条件下において時刻の読み取り間違いの数が記録された．観測値は Table 1 に示すとおりである．このデータを一元配置分散分析した結果の一部は Table 2 のようである．

Table 1

Dial type		
1	*2*	*3*
42	62	56
30	53	36
21	61	43
47	47	58
34	45	46
22	59	24
42		31
38		

Table 2

One-way ANOVA: type 1, type 2, type 3

```
Analysis of Variance
Source      DF         SS
Dial type    2       1377
Error       18       1858
Total       20       3234
```

(i) 分析を完全に行い，結果を解釈せよ．その際，結論に至るのに用いた仮定も示せ．

(ii) Type 1 および Type 2 の文字盤を読む際の誤りの平均個数の差の推定値およびその差の95%信頼区間を求めよ．そこで求めた「95%信頼区間」が何を意味するのかを述べよ．また，その解析が妥当であるための仮定を述べよ．

(iii) Table 1 のデータに関する一元配置分散分析に必要な仮定のチェックはどのように行えばよいかを述べよ．もしその仮定が満たされていないとした場合どのようにしたらよいかを述べよ（実際に実行する必要はない）．

4. 以下の各問に答えよ．

(i) 2 つの説明変数 x_1，x_2 がある目的変数 Y の予測のために用いられる．この分析で用いられる線形モデルを示し，そのモデルにおける各項の意味を示せ．

(ii) 以下のデータは幅 x_1 (cm) および高さ x_2 (cm) の二重窓の値段 Y (£) を表わしたものである．

x_1	66	183	124	239	66	109	196	251	165	81	249	142	170	254
x_2	122	122	122	122	30	58	61	76	86	81	117	61	30	30
Y	66	78	75	90	45	57	73	83	71	59	95	61	53	64

(a) x_1 に対する Y および x_2 に対する Y のプロットのグラフを描け．これらからどのような情報が得られるのかを簡潔に述べよ．

(b) Y および x_1，x_2 を用いた重回帰分析のコンピュータの出力は以下のようである．出力結果を統計の非専門家にも分かるように説明せよ．また，Y の x_1 および x_2 に対する回帰式を示せ．

```
Predictor      Coef     StDev        t        p
Constant     24.823     2.909     8.53    0.000
X1          0.13800   0.01288    10.72    0.000
X2          0.27226   0.02392    11.38    0.000

s = 3.132     Rsq = 95.9%

Analysis of variance

Source       df       SS       MS        F        p
Regression    2   2515.0   1257.5    128.2    0.000
Residual     11    107.9      9.8
Total        13   2622.9
```

(c) Y の x_1 に対する単回帰分析における回帰平方和は 1244.0 であった．このことから何がいえるのかを述べよ．