KHOA CÔNG NGHỆ THÔNG TIN
BỘ MÔN TOÁN

ĐỀ THI **KẾT THÚC** HỌC PHẦN

Tên Học phần: Phân tích số liệu
Thời gian làm bài: 60 phút
Loại đề thi: **Tự luận**

Ngày thi: 11/09/2020
Đề số: 01

Bài I (5,0 điểm) Đo nồng độ bụi trong không khí tại các khu vực trong một thành phố ở cùng một thời điểm, người ta thu được số liệu sau (đơn vị: mg / m^3):

Khu vực I	Khu vực II	Khu vực III	Khu vực IV
0.54	0.48	0.56	0.47
0.60	0.49	0.62	0.52
0.72	0.55	0.60	0.56
0.67	0.62	0.71	0.53
0.83	0.57	0.73	
0.63	0.59		

1) (1,0 đ) Lập mô hình phân tích phương sai một nhân tố cho bài toán trên.
2) (1,5 đ) Lập bảng phân tích phương sai cho bảng số liệu trên.
3) (1,0 đ) Tính nồng độ bụi trung bình trong không khí tại khu vực I và III.
4) (1,0 đ) Với mức ý nghĩa 5%, hãy xem có sự khác biệt về mức độ nhiễm bụi của bốn khu vực trên hay không?
5) (0,5 đ) Hãy sử dụng phương pháp Tukey HSD để kiểm định xem có sự khác biệt về mức độ nhiễm bụi của khu vực I và II ở mức ý nghĩa 0,05 hay không?

Bài II (5,0 điểm) Cho bảng số liệu sau:

y	x_1	x_2	x_3	x_4	x_5	x_6	x_7
36.98	5.1	400	51.37	4.24	1484.83	2227.25	2.06
13.74	26.4	400	72.33	30.87	289.94	434.9	1.33
10.08	23.8	400	71.44	33.01	320.79	481.19	0.97
8.53	46.4	400	79.15	44.61	164.76	247.14	0.62
36.42	7	450	80.47	33.84	1097.26	1645.89	0.22
26.59	12.6	450	89.9	41.26	605.06	907.59	0.76
19.07	18.9	450	91.48	41.88	405.37	608.05	1.71
5.96	30.2	450	98.6	70.79	253.7	380.55	3.93
15.52	53.8	450	98.05	66.82	142.27	213.4	1.97
56.61	5.6	400	55.69	8.92	1326.24	2043.36	5.08
26.72	15.1	400	66.29	17.98	507.65	761.48	0.6
20.8	20.3	400	58.94	17.79	377.6	566.4	0.9
6.99	48.4	400	74.74	33.94	158.05	237.08	0.63

45.93	5.8	425	63.71	11.95	130.66	1961.49	2.04
43.09	11.2	425	67.14	14.73	682.59	1023.89	1.57
15.79	27.9	425	77.65	34.49	274.2	411.3	2.38
21.6	5.1	450	67.22	14.48	1496.51	2244.77	0.32
35.19	11.7	450	81.48	29.69	652.43	987.64	0.44
26.14	16.7	450	83.88	26.33	458.42	687.62	8.82
8.6	24.8	450	89.38	37.98	312.25	468.38	0.02
11.63	24.9	450	79.77	25.66	307.08	460.62	1.72
9.59	39.5	450	87.93	22.36	193.61	290.42	1.88
4.42	29	450	79.5	31.52	155.96	233.95	1.43
38.89	5.5	460	72.73	17.86	1329.08	2088.12	1.35
11.19	11.5	450	77.88	25.2	663.09	994.63	1.61
75.62	5.2	470	75.5	8.66	1464.11	2196.17	4.78

Xét mô hình hồi qui tuyến tính sau:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \varepsilon \quad (1)$ với giả thiết $\varepsilon \sim N(0;\sigma^2)$.

1) (1,5 đ) Với mức ý nghĩa 0,05 hãy kiểm định xem có hay không mối quan hệ tuyến tính giữa biến phụ thuộc y với các biến giải thích $x_1, x_2, x_3, x_4, x_5, x_6, x_7$, tức là hãy kiểm định cặp giả thuyết, đối thuyết sau:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$$
$$H_1 : \exists \beta_i \neq 0,\ i \in \{1;2;3;4;5;6;7\}$$

2) (1,5 đ) Hãy ước lượng các hệ số hồi qui $\beta = (\beta_0;\beta_1;\beta_2;\beta_3;\beta_4;\beta_5;\beta_6;\beta_7)$. Từ đó hãy dự báo giá trị của y khi :

$$x_1 = 25,\ x_2 = 420,\ x_3 = 71.25,\ x_4 = 33.02,\ x_5 = 342.01,\ x_6 = 480.1,\ x_7 = 0.98$$

3) (0,75đ) Tính hệ số xác định đã hiệu chỉnh của mô hình.
4) (0,5 đ) Tính ước lượng không chệch của σ^2.
5) (0,75 đ) Hãy lựa chọn danh sách biến phù hợp để xây dựng mô hình tối ưu cho bài toán trên dựa vào giá trị AIC hoặc sử dụng thuật toán loại bỏ lùi (chọn $\alpha_{crit} = 0.2$).

..... Hết .....

Ghi chú: + Cán bộ coi thi không phải giải thích gì thêm.
+ Sinh viên được sử dụng tài liệu.

Cán bộ ra đề — *Nguyễn Hà Thanh*

Duyệt đề — *Vũ Thị Thu Giang*

KHOA CÔNG NGHỆ THÔNG TIN
BỘ MÔN TOÁN

ĐỀ THI **KẾT THÚC** HỌC PHẦN

Tên Học phần: Phân tích số liệu
Thời gian làm bài: 60 phút
Loại đề thi: **Tự luận**

Ngày thi: 11/09/2020
Đề số: 02

Bài I (5,0 điểm) Đo nồng độ bụi trong không khí tại các khu vực trong một thành phố ở cùng một thời điểm, người ta thu được số liệu sau (đơn vị: mg / m^3):

Khu vực I	Khu vực II	Khu vực III	Khu vực IV
0.53	0.47	0.58	0.45
0.61	0.49	0.61	0.52
0.72	0.56	0.60	0.57
0.67	0.62	0.70	0.55
0.82	0.58	0.72	
0.62	0.60		

1) (1,0 đ) Lập mô hình phân tích phương sai một nhân tố cho bài toán trên.
2) (1,5 đ) Lập bảng phân tích phương sai cho bảng số liệu trên.
3) (1,0 đ) Tính nồng độ bụi trung bình trong không khí tại khu vực II và IV.
4) (1,0 đ) Với mức ý nghĩa 5%, hãy xem có sự khác biệt về mức độ nhiễm bụi của bốn khu vực trên hay không?
5) (0,5 đ) Hãy sử dụng phương pháp Tukey HSD để kiểm định xem có sự khác biệt về mức độ nhiễm bụi của khu vực I và III ở mức ý nghĩa 0,05 hay không?

Bài II (5,0 điểm) Cho bảng số liệu sau:

y	x_1	x_2	x_3	x_4	x_5	x_6	x_7
36.89	5.1	400	51.37	4.24	1484.83	2227.15	2.06
13.74	26.4	400	72.33	30.87	289.94	434.9	1.33
10.08	23.8	400	71.44	33.01	320.79	481.19	0.97
8.53	46.4	400	79.15	44.61	164.76	247.14	0.62
36.42	7	420	80.47	33.84	1097.26	1645.24	0.22
26.59	12.6	420	89.9	41.26	605.06	907.59	0.76
19.07	18.9	420	91.48	41.88	405.37	608.05	1.71
5.96	30.2	420	98.6	70.79	253.7	380.55	3.93
15.52	53.8	420	98.05	66.82	142.27	213.4	1.97
55.89	5.6	400	55.69	8.92	1326.24	2042.26	5.08
26.72	15.1	400	66.29	17.98	507.65	761.48	0.6
20.8	20.3	400	58.94	17.79	377.6	566.4	0.9
6.99	48.4	400	74.74	33.94	158.05	237.08	0.63

45.93	5.8	415	63.71	11.95	130.66	1961.49	2.04
43.09	11.2	415	67.14	14.73	682.59	1023.89	1.57
15.79	27.9	415	77.65	34.49	274.2	411.3	2.38
21.6	5.1	420	67.22	14.48	1496.51	2243.15	0.32
35.19	11.7	420	81.48	29.69	652.43	987.64	0.44
26.14	16.7	420	83.88	26.33	458.42	687.62	8.82
8.6	24.8	420	89.38	37.98	312.25	468.38	0.02
11.63	24.9	420	79.77	25.66	307.08	460.62	1.72
9.59	39.5	420	87.93	22.36	193.61	290.42	1.88
4.42	29	420	79.5	31.52	155.96	233.95	1.43
38.89	5.5	430	72.73	17.86	1330.01	2078.12	1.35
11.19	11.5	420	77.88	25.2	663.09	994.63	1.61
75.62	5.2	450	75.5	8.66	1445.85	2196.17	4.78

Xét mô hình hồi qui tuyến tính sau:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \varepsilon \quad (1)$ với giả thiết $\varepsilon \sim N(0;\sigma^2)$.

1) (1,5 đ) Với mức ý nghĩa 0,05 hãy kiểm định xem có hay không mối quan hệ tuyến tính giữa biến phụ thuộc y với các biến giải thích $x_1, x_2, x_3, x_4, x_5, x_6, x_7$, tức là hãy kiểm định cặp giả thuyết, đối thuyết sau:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$$
$$H_1 : \exists \beta_i \neq 0,\ i \in \{1;2;3;4;5;6;7\}$$

2) (1,5 đ) Hãy ước lượng các hệ số hồi qui $\beta = (\beta_0;\beta_1;\beta_2;\beta_3;\beta_4;\beta_5;\beta_6;\beta_7)$. Từ đó hãy dự báo giá trị của y khi :

$x_1 = 25,\ x_2 = 400,\ x_3 = 71.25,\ x_4 = 32.02,\ x_5 = 342.01,\ x_6 = 479.1,\ x_7 = 0.98$.

3) (0,75đ) Tính hệ số xác định của mô hình.

4) (0,5 đ) Tính ước lượng không chệch của σ^2.

6) (0,75 đ) Hãy lựa chọn danh sách biến phù hợp để xây dựng mô hình tối ưu cho bài toán trên dựa vào giá trị AIC hoặc sử dụng thuật toán loại bỏ lùi (chọn $\alpha_{crit} = 0.2$).

..... Hết .....

Ghi chú: + Cán bộ coi thi không phải giải thích gì thêm.
+ Sinh viên được sử dụng tài liệu.

Cán bộ ra đề — *Nguyễn Hà Thanh*

Duyệt đề — *Vũ Thị Thu Giang*

KHOA CÔNG NGHỆ THÔNG TIN **BỘ MÔN TOÁN**	**ĐÁP ÁN ĐỀ THI KẾT THÚC HỌC PHẦN** **Tên học phần: phân tích số liệu** Đáp án đề thi số: 01

(Ngày thi: 11/09/2020)

Ghi chú : Mọi cách giải khác đáp án mà đúng đều được đủ điểm.

Câu		Đáp án vắn tắt	Điểm
I **5đ**	1	Mô hình: $x_{ij} = \mu_i + \varepsilon_{ij}$ x_{ij} là nồng độ bụi ở lần đo thứ j của khu vực i μ_i là nồng độ bụi trung bình của khu vực i ε_{ij} là sai số thực nghiệm của lần đo thứ j của khu vực i	1.0
	2	<table><tr><th>Nguồn KB</th><th>Bậc TD</th><th>TBP</th><th>PS</th><th>F</th><th>P_giá trị</th></tr><tr><td>Giữa các nhóm</td><td>3</td><td>0.07563</td><td>0.025208</td><td>4.644</td><td>0.0159</td></tr><tr><td>Trong nhóm</td><td>17</td><td>0.09227</td><td>0.005428</td><td></td><td></td></tr><tr><td>Tổng</td><td>20</td><td>0.1679</td><td></td><td></td><td></td></tr></table>	1.5
	3	$\overline{x_1} = 0.665; \overline{x_3} = 0.644$	0.5*2
	4	$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ H_1 : \exists i, j : \mu_i \neq \mu_j \end{cases}$ p _giá trị = 0.0159 < 0.05 nên bác bỏ H_0 Có sự khác biệt về nồng độ bụi ở các khu vực	0.25 0.5 0.25
	5	$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$ $p_{12} = 0.0653989 > 0.05$ nên chấp nhận H_0 Không có sự khác biệt về nồng độ bụi ở khu vực I và II	0.5
II **5đ**	1	$p_{gt} = 0.0005181 < 0.05$ nên bác bỏ H_0 Có mối quan hệ tuyến tính giữa biến phụ thuộc và các biến giải thích.	1.5
	2	Ước lượng của các hệ số hồi quy là $\begin{pmatrix} 54.805160 \\ -0.125198 \\ -0.232582 \\ 0.831240 \\ -0.438907 \\ -0.002027 \\ 0.020035 \\ 1.971933 \end{pmatrix}$	1.0
		Giá trị dự báo của y là 9.58194	0.5
	3	$R^2_{adj} = 0.6166$	0.75
	4	Ước lượng của σ^2 là 10.89	0.5
	5	Phương pháp loại bỏ lùi với $\alpha_{crit} = 0.2$ B1. Chạy mô hình đầy đủ được $p_5 = 0.8418 > 0.02$ và lớn nhất nên bỏ biến x_5	0.25
		B2. Chạy mô hình bỏ x_5 : $p_1 = 0.66654$ lớn nhất là >0.2 nên bỏ tiếp x_1 B3. Chạy mô hình bỏ x_5, x_1 : $p_2 = 0.35255$ lớn nhất là >0.2 nên bỏ tiếp x_2 B4. Chạy mô hình bỏ x_5, x_1, x_2 : $p_3 = 0.6280$ lớn nhất là >0.2 nên bỏ tiếp x_3 B5. Chạy mô hình bỏ x_5, x_1, x_2, x_3 : $p_4 = 0.41224$ lớn nhất là >0.2 nên bỏ tiếp x_4	0.25
		B6. Chạy mô hình bỏ x_5, x_1, x_2, x_3, x_4 : $p_6 = 9.6 \times 10^{-7}, p_7 = 0.0559$ đều < 0.2 nên chọn mô hình này: $y = \beta_0 + \beta_6 x_6 + \beta_7 x_7 + \varepsilon$	0.25
		AIC = 123,2 Mô hình tối ưu gồm các biến x_6, x_7 $y = \beta_0 + \beta_6 x_6 + \beta_7 x_7 + \varepsilon$	0.75

Cán bộ ra đề
Nguyễn Hà Thanh
CB làm đáp án
Lê Thị Diệu Thùy

Duyệt đề
Nguyễn Thị Bích Thủy

KHOA CÔNG NGHỆ THÔNG TIN **BỘ MÔN TOÁN**	**ĐÁP ÁN ĐỀ THI KẾT THÚC HỌC PHẦN** **Tên học phần: phân tích số liệu** Đáp án đề thi số: 02

(Ngày thi: 11/09/2020)

Ghi chú : Mọi cách giải khác đáp án mà đúng đều được đủ điểm.

<table>
<tr><th colspan="2">Câu</th><th>Đáp án vắn tắt</th><th>Điểm</th></tr>
<tr><td rowspan="5">I
5đ</td><td>1</td><td>Mô hình: $x_{ij} = \mu_i + \varepsilon_{ij}$
x_{ij} là nồng độ bụi ở lần đo thứ j của khu vực i
μ_i là nồng độ bụi trung bình của khu vực i
ε_{ij} là sai số thực nghiệm của lần đo thứ j của khu vực i</td><td>1.0</td></tr>
<tr><td>2</td><td><table><tr><th>Nguồn KB</th><th>Bậc TD</th><th>TBP</th><th>PS</th><th>F</th><th>P_giá trị</th></tr><tr><td>Giữa các nhóm</td><td>3</td><td>0.06872</td><td>0.022908</td><td>4.189</td><td>0.0216</td></tr><tr><td>Trong nhóm</td><td>17</td><td>0.09297</td><td>0.005469</td><td></td><td></td></tr><tr><td>Tổng</td><td>20</td><td>0.16169</td><td></td><td></td><td></td></tr></table></td><td>1.5</td></tr>
<tr><td>3</td><td>$\overline{x_2} = 0.5533333; \overline{x_4} = 0.5225$</td><td>0.5*2</td></tr>
<tr><td>4</td><td>$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ H_1 : \exists i, j : \mu_i \neq \mu_j \end{cases}$
p _giá trị = 0.0216 < 0.05 nên bác bỏ H_0
Có sự khác biệt về nồng độ bụi ở các khu vực</td><td>0.25
0.5
0.25</td></tr>
<tr><td>5</td><td>$\begin{cases} H_0 : \mu_1 = \mu_3 \\ H_1 : \mu_1 \neq \mu_3 \end{cases}$
$p_{13} = 0.9708093 > 0.05$ nên chấp nhận H_0
Không có sự khác biệt về nồng độ bụi ở khu vực I và III</td><td>0.5</td></tr>
<tr><td>II
5đ</td><td>1</td><td>$p_{gt} = 0.0004346 < 0.05$ nên bác bỏ H_0
Có mối quan hệ tuyến tính giữa biến phụ thuộc và các biến giải thích.</td><td>1.5</td></tr>
<tr><td></td><td>2</td><td>Ước lượng của các hệ số hồi quy là $\begin{pmatrix} -142.602 \\ -0.06932 \\ 0.4574 \\ -0.4817 \\ 0.03740 \\ -0.003324 \\ 0.01338 \\ 1.884 \end{pmatrix}$
Giá trị dự báo của y là 12.64045</td><td>1.0
0.5</td></tr>
<tr><td rowspan="4"></td><td>3</td><td>$R^2_{adj} = 0.6248$</td><td>0.75</td></tr>
<tr><td>4</td><td>Ước lượng của σ^2 là 10.74</td><td>0.5</td></tr>
<tr><td rowspan="3">5</td><td>Phương pháp loại bỏ lùi với $\alpha_{crit} = 0.2$
B1. Chạy mô hình đầy đủ được $p_4 = 0.918 > 0.2$ và lớn nhất nên bỏ biến x_4</td><td>0.25</td></tr>
<tr><td>B2. Chạy mô hình bỏ x_4:
$p_1 = 0.817$ lớn nhất và >0.2 nên bỏ tiếp x_1
B3. Chạy mô hình bỏ x_4, x_1:
$p_5 = 0.7352$ lớn nhất và >0.2 nên bỏ tiếp x_5</td><td>0.25</td></tr>
<tr><td>B4. Chạy mô hình bỏ x_4, x_1, x_5
$p_2 = 0.131, p_3 = 0.1795, p_6 = 0.0188, p_7 = 0.0973$
đều < 0.2 nên chọn mô hình này:
$y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_6 x_6 + \beta_7 x_7 + \varepsilon$</td><td>0.25</td></tr>
<tr><td></td><td>AIC = 124,14.
Mô hình tối ưu gồm các biến x_2, x_3, x_6, x_7
$y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_6 x_6 + \beta_7 x_7 + \varepsilon$</td><td>0.75</td></tr>
</table>

Cán bộ ra đề
Nguyễn Hà Thanh
CB làm đáp án
Lê Thị Diệu Thùy

Duyệt đề
Nguyễn Thị Bích Thủy