The Hong Kong University of Science & Technology

MATH3424 - Regression Analysis
Final Examination

Answer ALL Questions **Date: 16 December 2020**

Full marks: 70 + Bonus: 4 **Time Allowed: 3 hours**

---

1. **(12 marks)** Consider a linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + e_i$$
$$\text{for } i = 1, 2, ..., 52 \ \& \ e_i \sim i.i.d.N(0, \sigma^2)$$

Use the table below to answer the following questions

Model (β_0 included)	Res.SS	Model (β_0 included)	Res.SS
none	12313	x_1, x_2, x_3	9033.92
x_1	11257	x_1, x_2, x_4	6806.40
x_2	9689.95	x_1, x_2, x_5	7992.67
x_3	12224	x_1, x_3, x_4	6395.31
x_4	8065.39	x_1, x_3, x_5	9033.50
x_5	9246.53	x_1, x_4, x_5	6455.48
x_1, x_2	9267.16	x_2, x_3, x_4	6713.73
x_1, x_3	11189	x_2, x_3, x_5	7793.58
x_1, x_4	6941.03	x_2, x_4, x_5	6758.55
x_1, x_5	9033.71	x_3, x_4, x_5	6511.71
x_2, x_3	9396.06	x_1, x_2, x_3, x_4	5980.03
x_2, x_4	7626.16	x_1, x_2, x_3, x_5	7727.32
x_2, x_5	8088.52	x_1, x_2, x_4, x_5	6364.19
x_3, x_4	7594.91	x_1, x_3, x_4, x_5	5932.15
x_3, x_5	9246.51	x_2, x_3, x_4, x_5	5964.02
x_4, x_5	6978.89	x_1, x_2, x_3, x_4, x_5	5602.49

(a) **(4 marks)** Find the best model by stepwise regression. Write down how to get the best model in details. Choose critical values for both ENTRY and STAY to be 4. Which variables are in the best model?

(b) **(3 marks)** Find the smallest value of C_p for $p = 1, \ldots, 5$, where p is the number of independent variables. Hence, find the best model based on these C_p values. Which variables are in the best model?

(c) **(3 marks)** Compute R^2, AIC and BIC for the models obtained from parts (a) and (b). Which model do you recommend? Why?

(d) **(2 marks)** Write down the test statistic for testing the significance of regression in terms of R^2 (coefficient of determination), n (sample size) and p (the number of independent variables).

2. **(19 marks)** A sample of elderly people were given a psychiatric examination to determine whether symptoms of senility are present. One explanatory variable is the score on a subtest of the Wechsler Adult Intelligence Scale. Consider a logistic model for the probability of having symptoms of senility on WAIS score.

   (a) The table below shows the summary of the maximum likelihood estimates and their variance and covariance matrix.

Parameter	Estimate	Covariance Matrix	
		Intercept	WAIS score
Intercept	2.4040	1.420471	-0.12997
WAIS score	-0.3235	-0.12997	0.012991

   Based on the above table, answer the following questions.

      i. **(1 mark)** Write down the fitted model.

      ii. **(2 marks)** Estimate the odds ratio when WAIS score is increased by one unit. Hence or otherwise, find the percentage increase (or reduction) of odds when WAIS score is increased by one unit.

      iii. **(1 mark)** For which region of WAIS scores does the predicted probability of the presence of symptoms exceed 0.5?

      iv. **(5 marks)** Estimate the probability (with 95% confidence interval) of having symptoms of senility on WAIS score at score = 10. Then, test whether this probability is significantly different from 0.5 by Wald test at $\alpha = 0.05$?

   (b) Observations are then separated into two groups according to their WAIS score. It is noted that 4 (out of 32 elderly people with WAIS score ≥ 10) and 10 (out of 22 elderly people with WAIS score < 10) have symptoms of senility. Consider a logistic model for the probability of having symptoms of senility on WAIS score as a categorical variable (with score less than 10 as reference group).

      i. **(1 mark)** Write down the likelihood function of unknown parameters, β_0 and β_1, for obtaining maximum likelihood estimates.

      ii. **(2 marks)** Find the fitted model.

      iii. **(3 marks)** Estimate the odds ratio (with 95% confidence interval) of having symptoms of senility for an elderly people with WAIS score ≥ 10 vs an elderly people with WAIS score < 10.

      iv. **(4 marks)** Estimate the probability (with 95% confidence interval) of having symptoms of senility for an elderly people with score ≥ 10.

3. **(29 marks)** An experiment was conducted to study the effect of temperature and type of oven on the lift of a particular component being tested. Two temperature levels and two types of ovens were used in the experiment. The following summary statistics on the lift of a particular component being tested (y) were recorded:

Temp	Oven								
	1			2			Total		
	n	Sum	UCSS	n	Sum	UCSS	n	Sum	UCSS
1	4	921	213247	6	1469	361407	10	2390	574654
2	8	1523	290955	6	1246	260478	14	2769	551433
Total	12	2444	504202	12	2715	621885	24	5159	1126087

where UCSS is uncorrected sum of squares, i.e., UCSS $= \sum y_i^2$.

(a) **(2 marks)** Consider a homogeneity model, i.e.,

$$\boldsymbol{y_i} \;=\; \boldsymbol{\beta_0} + \boldsymbol{e_i}$$

for $i = 1, \ldots, 24$, where $e_i \sim N(0, \sigma^2)$. Find the unbiased estimate of the unknown parameter σ^2. No need to show that it is unbiased.

(b) Consider a model of y on temperature (defined as x_1 with 2 as reference group), i.e.,

$$\boldsymbol{y_i} \;= \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_{i1}} + \boldsymbol{e_i}$$

for $i = 1, \ldots, 24$, where $x_{i1} = 0$ if level of temperature $= 2$ and $e_i \sim N(0, \sigma^2)$.

i. **(2 marks)** Find the unbiased estimate of the unknown parameter σ^2. No need to show that it is unbiased.

ii. **(3 marks)** Test the null hypothesis that population means of y for two levels of temperature are equal at the 5% significance level. Write down the null hypothesis, test statistic, critical value and your conclusion clearly.

(c) Define μ_{ij} to be the population mean of y with i for level of temperature and j for type of oven where $i, j = 1, 2$. Consider a model of y on the population means of these four parts.

i. **(2 marks)** Estimate all unknown parameters in the model.

ii. **(2 marks)** Find the unbiased estimate of the unknown parameter σ^2. No need to show that it is unbiased.

iii. **(4 marks)** Test the population means of these four parts are equal at the 5% significance level. Write down the null hypothesis, test statistic, critical value and your conclusion clearly.

(d) Consider a model of y on temperature (defined as x_1 with 2 as reference group) and oven (defined as x_2 with 2 as reference group) and their interaction term, i.e.,

$$\boldsymbol{y_i} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_{i1}} + \boldsymbol{\beta_2 x_{i2}} + \boldsymbol{\gamma x_{i1} x_{i2}} + \boldsymbol{e_i}$$

for $i = 1, \ldots, 24$, where $x_{i1} = 0$ if level of temperature $= 2$, $x_{i2} = 0$ if type of oven $= 2$ and $e_i \sim N(0, \sigma^2)$.

i. **(2 marks)** Estimate the unknown parameters in the model.

ii. **(1 mark)** Find the unbiased estimate of the unknown parameter σ^2. No need to show that it is unbiased.

iii. **(4 marks)** Calculate the sum of squares for the interaction term of "x_1" and "x_2". Then, test the null hypothesis that there is no interaction between x_1 and x_2 at the 5% significance level by <u>F-test</u>. Write down the null hypothesis, test statistic, critical value and your conclusion clearly.
Hint: Write down the null hypothesis that there is no interaction between x_1 and x_2 as $H_0 : \underset{\sim}{C}\underset{\sim}{\beta} = \underset{\sim}{d}$, where $\underset{\sim}{\beta} = (\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22})^T$.

(e) Assume that the interaction between temperature (x_1) and oven (x_2) is insignificant, i.e.,

$$\boldsymbol{y_i} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_{i1}} + \boldsymbol{\beta_2 x_{i2}} + \boldsymbol{e_i}$$

for $i = 1, \ldots, 24$, where $x_{i1} = 0$ if level of temperature $= 2$, $x_{i2} = 0$ if type of oven $= 2$ and $e_i \sim N(0, \sigma^2)$.

i. **(2 marks)** Find the unbiased estimate of the unknown parameter σ^2. No need to show that it is unbiased.

ii. **(1 mark)** Find the Regression Sum of Squares.

iii. **(4 marks)** Hence or otherwise, find the sum of squares for the main effect of oven on y. Then, test the null hypothesis that the population means of y for two types of oven are equal at the 5% significance level. Write down the null hypothesis, test statistic, the critical value and your conclusion clearly.

4. **(10 marks + Bonus: 4 marks)** In a project, age and growth characteristics of selected mussel species from Southwestern Virginia in two locations were studied. Sample size, Res.S.S., summary statistics of weight (y), parameter estimates, standard error and covariance matrix of regression with weight as the response and age as the independent variable for each location are given below:

<u>Location=1</u>

$n = 35$, Res.S.S. $= 25.25034$, $\sum_{i=1}^{35} y_i = 111.56$, $\sum_{i=1}^{35} y_i^2 = 476.9816$, $S_{yy} = 121.392069$,

			Covariance Matrix	
Variable	$\hat{\beta}_i$	St. Error	Intercept	age
Intercept	-0.48283	0.35927	0.129071	-0.0106601
age	0.36494	0.03256	-0.0106601	0.00105995

<u>Location=2</u>

$n = 30$, Res.S.S. $= 22.99603$, $\sum_{i=1}^{35} y_i = 161.18$, $\sum_{i=1}^{35} y_i^2 = 1252.13$, $S_{yy} = 386.1643867$,

			Covariance Matrix	
Variable	$\hat{\beta}_i$	St. Error	Intercept	age
Intercept	-1.24204	0.35542	0.126324	-0.00979685
age	0.65492	0.03114	-0.00979685	0.000969986

Under the assumption that the population variances of weight for both locations are equal, we consider to fit a regression line with weight as the response (y), age as the independent variable (x_1), and location as a categorical variable with 2 as reference group (i.e., $x_2 = 1$ for Location = 1; $x_2 = 0$ for Location = 2), i.e.,

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2} + e_i}$$

for $i = 1 \ldots 65$.

(a) **(2 marks)** Find the fitted line for the model of y on x_1, x_2 and their interaction term(s).

(b) **(2 marks)** Find an unbiased estimate of the common variance, which is more efficient than other unbiased estimates from the data set. Explain why it is more efficient.

(c) **(2 marks)** Calculate the Total Sum of Squares (Total S.S.). Then, find the multiple correlation coefficient.

(d) **(4 marks)** Test $\beta_{12} = 0$ against the alternative hypothesis that $H_1 : \beta_{12} \neq 0$ by <u>t-test</u> at the 5% significance level. Write down the null hypothesis, test statistic, critical value and your conclusion clearly. Is one regression slope an appropriate mode?

(e) **(Bonus: 4 marks)** **Assume that $\boldsymbol{\beta_{12} \neq 0}$.** Estimate the difference of $E(y)$ between $x_{1c} = 1$ and $x_{1c} = 0$ at $x_2 = 10$ and then test whether it is significant at $\alpha = 0.05$ by <u>t-test</u>. Write down the test statistic, critical value and your decision rule clearly.

********** END **********