UNIVERSITY OF TORONTO
FACULTY OF APPLIED SCIENCE AND ENGINEERING

FINAL EXAMINATION, April, 2014 WITH SOLUTIONS

Second Year — Industrial and Mechanical

MIE237H1 — Statistics

Calculator Type: 2

Exam Type: B

Examiner: N. Montgomery

a) **An asterisk following a solution indicates that the solution given is factual only and would not have been considered a complete solution.**

b) There are 13 pages including this page.

c) There are 50 marks in total. The number of marks available for each part of each question is indicated.

FAMILY NAME:_____

GIVEN NAME:_____

STUDENT NUMBER:_____

The Story

A processed food packaging plant operates as follows. Raw materials are shipped to the plant from a variety of sources, generally consisting of large sacks of dried foods, flavorings, and the packaging materials and package labels themselves. The plant assembles the raw materials, puts them into packages, labels the packages, and organizes the packages into boxes for shipping to distributors and stores. Much of the work is automated, but the production lines do have to be operated and maintained by employees.

The plant hires an industrial engineer (you) who is good at statistics (because you listened to me so carefully) to evaluate various aspects of plant operations.

Most of this exam will involve answering questions about these evaluations, with a little theory thrown in here and there.

1. **(10 marks total)** The plant has 8 main production lines numbered 1 to 8. The current procedure for maintaining the lines is to have two technicians look after lines 1, 2, 3, and 4, and another two technicians to look after lines 5, 6, 7, and 8. If one of the groups of four lines has more than two problems that need fixing, then the line just remains shut down until one of the two technicians becomes available.

The plant proposes that it might be more efficient to have all four technicians look after all eight lines, with the idea being that while the technicians might have to spend more time getting to each production line on average, it would be less likely to have lines waiting for an available technician.

You decide to write a computer simulation that will compare the current procedure with the proposed procedure. The simulation considers technician travel times, failure rates, repair rates, and several other important factors. The key output variable being measured is "total minutes of production line downtime".

a) **(3 marks)** You run the simulation 60 times. 30 times using the current procedure specification, and 30 times using the proposed specification. Each run simulates one year of operating time. Each run is independent of all other runs. A summary of the results (total minutes of downtime) is contained in the following table:

	Sample Average	Sample Standard Deviation
Current	25731	520
Proposed	25399	361

Perform the hypothesis test to evaluate the evidence against the null hypothesis "there is no difference between the two procedures", using a p-value in your conclusion.

```
Difference = mu (1) - mu (2)
Estimate for difference:  322
95% CI for difference:  (91, 553)
T-Test of difference = 0 (vs not =): T-Value = 2.79  P-Value = 0.007  DF = 58
Both use Pooled StDev = 447.6165
```


***Indicates incomplete answer**

b) **(4 marks)** You have been given enough information in this question to evaluate the extent to which one, but not both, of the model assumptions have been satisfied for the conclusion in a) to be valid. Evaluate the assumption you can. Also, give an argument as to why the other assumption is very likely also sufficiently satisfied in this case.

Sample variance ratio is less than 3:1. Normality is likely satisfied since each simulation is affected by a moderately large number of factors, and also there is a moderately large sample size for each group.

c) **(3 marks)** In theory, the data could also have been analyzed using a paired t-test, say by numbering the runs for each procedure from 1 to 30, treating them as pairs, and then examining the 30 paired differences using a one sample t-test. State whether the resulting p-value would *likely* be larger, smaller, or the same as the one you obtained in a), and briefly state why.

It would likely be larger due to the lower degrees of freedom for the paired test.*

***Indicates incomplete answer**

2. **(25 marks total)** One of the products the plant makes is supposed to have a gross weight (including packaging) of 128 grams. The following variables are continuously measured and you wonder what the impact of these variables might be on the gross weight of one item of this product:

x_1**:** plant relative humidity in percent;

x_2**:** plant temperature in degrees Celsius;

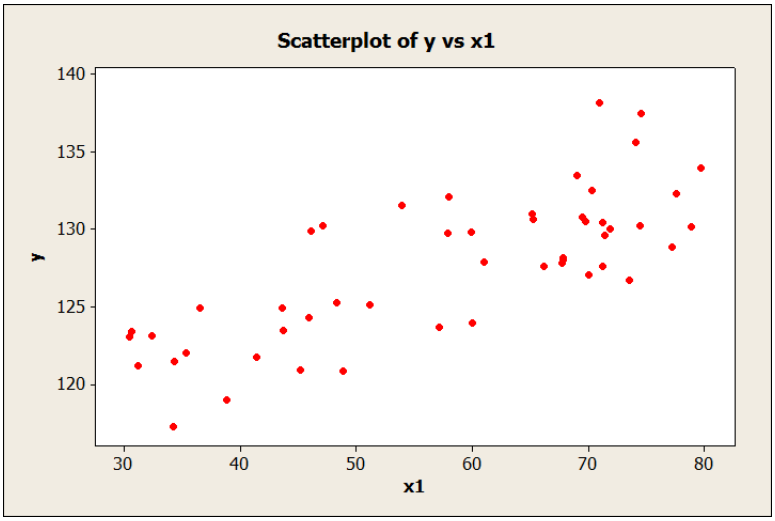
x_3**:** production line speed in packages produced per minute;

x_4**:** coded 0 for "Salt Supplier A" or 1 for "Salt Supplier B".

Each day for 50 days you take one finished item from the production line and measure its gross weight, and you also take note of values of each of the four variables listed above at the moment the item is taken from the line, resulting in a spreadsheet with 5 columns and 50 rows of data.

The sample variance $S_{yy}/(50-1)$ of the gross weights is 22.035.

a) **(8 marks)** Consider the simple regression model with y (gross weight) and x_1 (relative humidity). Here is a plot of the data:



***Indicates incomplete answer**

On the previous page you were told that the sample variance $S_{yy}/(50-1)$ of the gross weights is 22.035. In addition, the sample variance $S_{x_1x_1}/(50-1)$ of the relative humidity is 236.30. When the simple linear regression model of y on x_1 is run, you get SSR= 641.00.

Here is some Minitab output of the regression with 16 entries missing. Fill in all the missing entries.

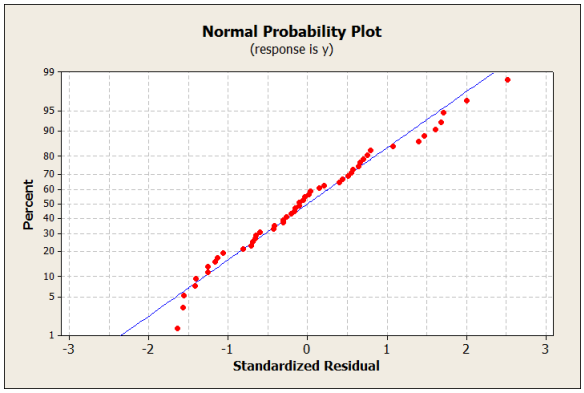
```
Predictor     Coef  SE Coef      T      P
Constant   113.992    1.679  67.88  0.000
x1         _______  _______  _____  _____

S = _______   R-Sq = _____

Analysis of Variance

Source          DF       SS      MS      F      P
Regression      ___  ______  ______  _____  _____
Residual Error  ___  ______  ______
Total           ___  ______

Unusual Observations

Obs    x1        y      Fit  SE Fit  Residual  St Resid
  4  74.5  137.454  131.531   0.636     5.922      2.00R
 28  71.0  138.174  130.688   0.565     7.487      2.52R

R denotes an observation with a large standardized residual.
```

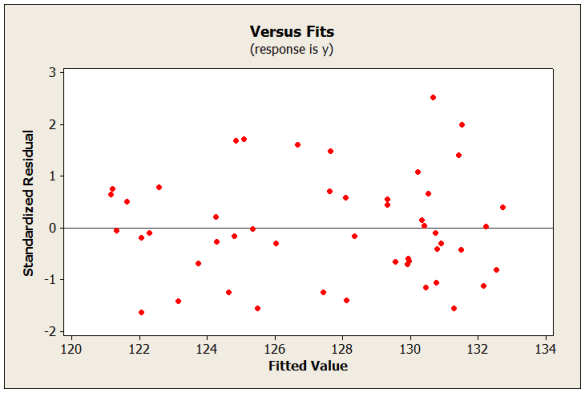
Answer:

```
Predictor     Coef  SE Coef      T      P
Constant   113.992    1.679  67.88  0.000
x1         0.23529  0.02810   8.37  0.000

S = 3.02314   R-Sq = 59.4%   R-Sq(adj) = 58.5%

Analysis of Variance

Source          DF       SS      MS      F      P
Regression       1   641.00  641.00  70.14  0.000
Residual Error  48   438.69    9.14
Total           49  1079.69
```


***Indicates incomplete answer**

b) **(3 marks)** Here is the normal quantile plot of the standardized residuals and the plot of standardized residuals versus fitted values.





Comment on the extent to which the usual simple regression model assumptions have been satisfied.

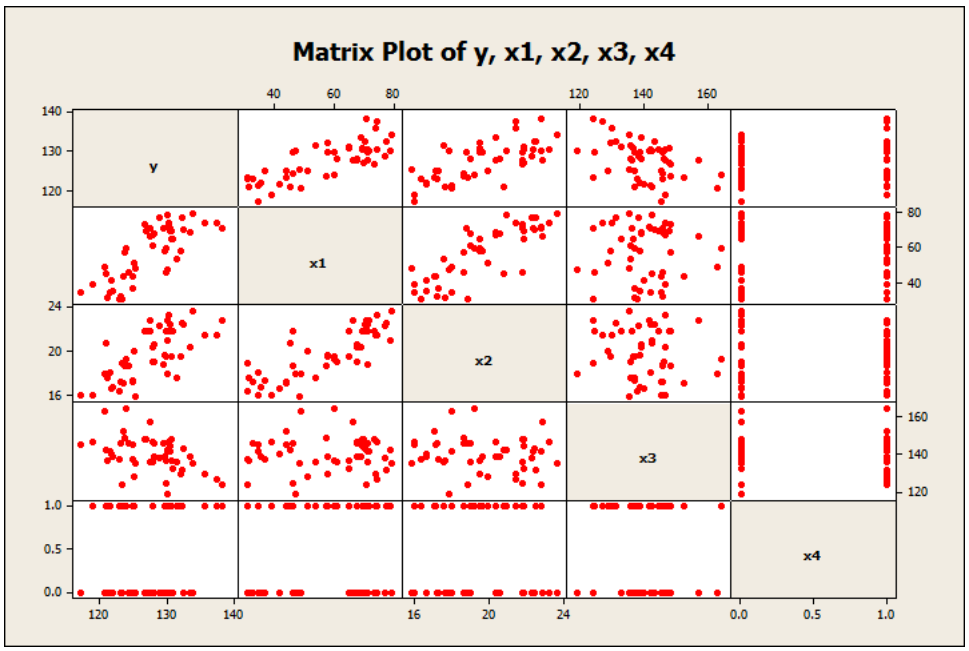
All have been satisfied.*

c) **(3 marks)** The sample average relative humidity is $\bar{x}_1 = 57.80\%$. Produce a 99% interval, if it is possible to do so accurately, for the gross weight of an item packaged when the relative humidity is 60%. If not, state why not.

(119.918, 136.300)

***Indicates incomplete answer**

d) **(3 marks)** Now consider all the variables. Here is a matrix scatterplot for the output y and all the inputs x_1, x_2, x_3, x_4.



Which variables are likely to be included in a final multiple regression model, and which variables are likely to cause problems in the multiple regression model selection process?

Included: x_1, x_2, x_3. Problem: x_1, x_2.*

***Indicates incomplete answer**

e) **(4 marks)** When you consider all the variables, there are 15 possible linear regression models of the type $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$ with $\varepsilon \sim N(0, \sigma^2)$. Excluding an x_i from a model is done by setting its β_i to 0.

Modeling results for all 15 models are summarized in the following table.

The "p-values" entries correspond to p-values for the hypothesis tests $H_0 : \beta_i = 0$ versus $H_1 : \beta_i \neq 0$ for all the β_i included in that row's model, correct to four decimal places. If a p-value entry is 0.0000, it means that p-value is less than 0.00005. If the p-value entry under β_i is '-', it means β_i was set to 0 so that x_i wasn't included in that model. The R^2 for each model is also included, expressed as a percentage.

Number of Inputs	Model #	p-values				R^2
		β_1	β_2	β_3	β_4	
1	1	0.0000	-	-	-	59.4
	2	-	0.0000	-	-	48.2
	3	-	-	0.0013	-	19.6
	4	-	-	-	0.2259	3.0
2	5	0.0004	0.2462	-	-	60.5
	6	0.0000	-	0.0000	-	80.6
	7	0.0000	-	-	0.1302	61.3
	8	-	0.0000	0.0000	-	62.3
	9	-	0.0000	-	0.1242	50.8
	10	-	-	0.0019	0.3246	21.3
3	11	0.0000	0.8552	0.0000	-	80.6
	12	0.0004	0.2218	-	0.1199	62.6
	13	0.0000	-	0.0000	0.1559	81.4
	14	-	0.0000	0.0002	0.1749	63.8
4	15	0.0000	0.7952	0.0000	0.1569	81.5

For example, Model 12 includes x_1, x_2, and x_4, but not x_3 because the p-value for β_3 is given as '-'.

Using this table, determine a good multiple regression model, and for the model you have chosen, perform the general hypothesis test that answers the question "is there any linear relationship between the output variable and the input variables", being careful to specify the null and alternative hypotheses and using a p-value in your conclusion.

(There is space on the next page to answer.)

***Indicates incomplete answer**

(This page is for answering question 2.e))

Model 6*

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	870.27	435.14	97.66	0.000
Residual Error	47	209.42	4.46		
Total	49	1079.69			

***Indicates incomplete answer**

f) **(4 marks)** (Theory) You are sitting in the computer lab beside a friend that you secretly have a huge crush on. You are both using Minitab (the world's most romantic software) to analyze the same bivariate dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ using the usual simple linear regression model. Your friend fits the model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with $\varepsilon_i \sim N(0, \sigma^2)$ and announces getting a p-value of 0.042 and "jokes" that if you get the same p-value then you will get a big kiss.

You are so nervous that you accidentally fit the model $x_i = \beta_0 + \beta_1 y_i + \varepsilon_i$, i.e. you reverse x and y.

Prove that despite your error you will still get the same p-value as your friend.

(Hint: first show that in any simple regression model it is always true that the $F = \frac{MSR}{MSE} = \frac{R^2(n-2)}{1-R^2}$)

Use the hint, plus the fact that $r^2 = R^2$ and the correlation coefficient is symmetric in x and y*

***Indicates incomplete answer**

3. **(10 marks total)** The food processing plant management asks you to compare the productivity of the eight processing lines, measured as the number of items produced per minute.

Lines 7 and 8 are a few years older than lines 1 to 6, so they are interested in seeing if there is any difference between the average of 7 and 8 versus the average of 1 to 6.

So you select 5 days at random from the next month and measure the number of items produced per minute for each line on each of the five days. The following table summarizes the results:

Level	N	Mean	StDev
1	5	191.15	7.48
2	5	188.02	16.45
3	5	202.79	6.66
4	5	205.91	14.38
5	5	181.30	9.16
6	5	207.82	7.04
7	5	204.25	10.69
8	5	198.84	5.87

In this question you may assume that any required model assumptions are satisfied.

a) **(4 marks)** The total sum of squares is 6665. The error sum of squares is 3439. Perform the hypothesis test that answers the question "is there any difference among the production lines?"

Source	DF	SS	MS	F	P
lines	7	3226	461	4.29	0.002
Error	32	3439	107		
Total	39	6665			

***Indicates incomplete answer**

b) **(3 marks)** Perform the hypothesis test that answers the question concerning lines 7 and 8 versus the other lines that was stated above.

$\hat{\omega} = 5.38$, $\widehat{\mathrm{Var}}(\hat{\omega}) = 14.267$, $T = 1.42$, $p = 0.164$

c) **(3 marks)** Determine the extent to which there are any differences between any pairs of production lines, at the $\alpha = 0.05$ experimentwise error rate.

F test significant at α level, so proceed...

lines	N	Mean	Grouping
6	5	207.82	A
4	5	205.91	A
7	5	204.25	A
3	5	202.79	A
8	5	198.84	A B
1	5	191.15	A B
2	5	188.02	A B
5	5	181.30	B

***Indicates incomplete answer**

4. **(5 marks total)** The plant operates two 8 hour shifts per day, called the day shift and the evening shift. Each shift is divided into two halves. The early half of the shift is followed by a meal break and then the second half. The plant is interested in the effect of day versus evening shift and first versus second half of shift on the number of production line faults, to see if fatigue or boredom is affecting reliability.

You go back into the production line fault records and determine that there were 412 faults in total during the past six months. 87 of them occurred in the first half of day shifts. 99 occurred in the second half of days shifts. 114 occurred in the first half of evening shifts. 112 occurred in the second half of evening shifts.

Perform the hypothesis test to determine if shift timing (day/evening) and shift half (early/late) are independent.

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

           C1      C2  Total
    1      87      99    186
        90.74   95.26
        0.154   0.147

    2     114     112    226
       110.26  115.74
        0.127   0.121

Total     201     211    412

Chi-Sq = 0.549, DF = 1, P-Value = 0.459
```


***Indicates incomplete answer**