UNIVERSITY OF TORONTO
FACULTY OF APPLIED SCIENCE AND ENGINEERING

FINAL EXAMINATION, April, 2013

Second Year — Industrial and Mechanical

MIE237H1 — Statistics & Design of Experiments

Calculator Type: 2

Exam Type: B

Examiner: N. Montgomery

a) There are 14 pages including this page. Do not detach pages from the exam paper itself.

b) You should also have an 4 page package printed on two sides. The first two sides are the aid sheet. The remaining six sides are tables of probabilities and critical values.

c) You can use the backs of the question pages for rough work.

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

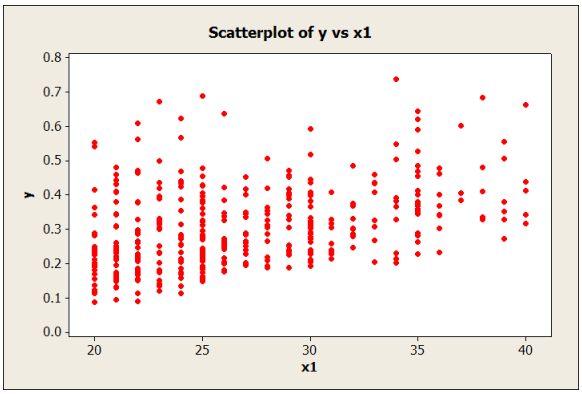
e) You will not be evaluated on the quality of your writing. Do not waste time writing lengthy answers when concise ones will do just as well.

FAMILY NAME:_____

GIVEN NAME:_____

STUDENT NUMBER:_____

1. **(10 marks total)** A natural gas distribution company takes a sample of $n = 400$ copper pipes that are used to supply gas to individual houses. The pipes range in age from 20 to 40 years, where two pipes installed during the same calendar year are considered to be the same age. The pipes corrode over time, which can lead to gas leaks. The measurement of interest y is the minimum wall thickness of the pipe in millimeters. Here is a plot of y_i versus the age of the pipe x_{1i} for the sample. (Age has been called x_1 because more variables x_2 and x_3 will be considered in a subsequent question.)

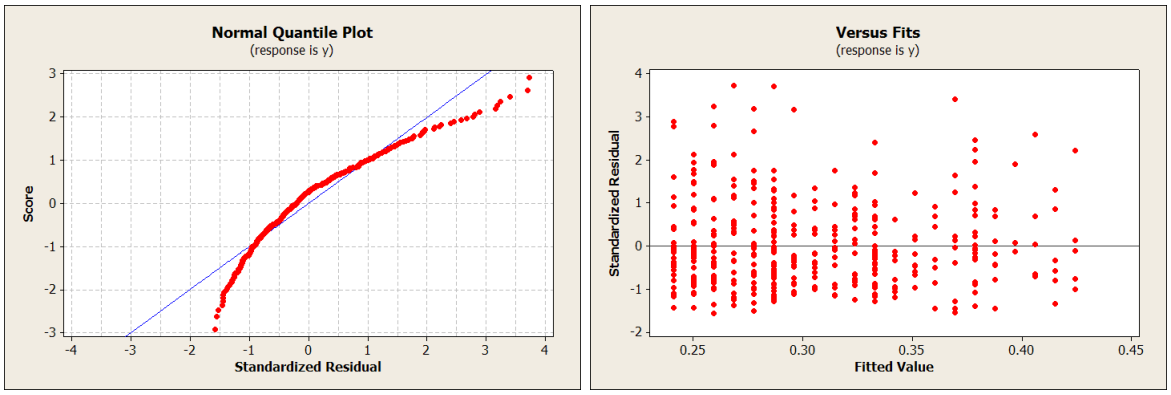


Here is the Minitab output with some numbers removed:

```
Predictor       Coef    SE Coef      T      P
Constant     0.05801    0.02785   2.08  0.038
x1           *******    *******   ****  *****


S = 0.108327   R-Sq = *****


Analysis of Variance

Source            DF       SS       MS      F      P
Regression       ***  0.93948  *******  *****  *****
Residual Error   ***  *******  *******
Total            ***  5.60990
```

a) **(3 marks)** Perform the hypothesis test $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ using a p-value in your conclusion.

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



Comment on the extent to which the usual simple regression model assumptions are satisfied, and if your conclusion in 1.a) is valid.

c) **(2 marks)** Here is the list of unusual observations that Minitab produces (it extends to the next page!)

Unusual Observations

Obs	x	y	Fit	SE Fit	Residual	St Resid
15	22.0	0.56302	0.25961	0.00723	0.30341	2.81R
27	20.0	0.55403	0.24128	0.00872	0.31275	2.90R
55	21.0	0.47964	0.25044	0.00794	0.22919	2.12R
68	22.0	0.60993	0.25961	0.00723	0.35032	3.24R
82	24.0	0.62288	0.27793	0.00607	0.34494	3.19R
94	23.0	0.67200	0.26877	0.00659	0.40323	3.73R
104	24.0	0.56702	0.27793	0.00607	0.28909	2.67R
133	20.0	0.54226	0.24128	0.00872	0.30098	2.79R
162	25.0	0.68837	0.28710	0.00568	0.40127	3.71R
182	23.0	0.49914	0.26877	0.00659	0.23037	2.13R
297	26.0	0.63862	0.29626	0.00546	0.34236	3.16R
310	30.0	0.59373	0.33291	0.00640	0.26081	2.41R
343	34.0	0.73766	0.36957	0.00926	0.36809	3.41R
369	39.0	0.50719	0.41539	0.01374	0.09180	0.85 X
371	40.0	0.66244	0.42455	0.01469	0.23789	2.22RX
374	38.0	0.68439	0.40622	0.01281	0.27817	2.59R
375	39.0	0.55639	0.41539	0.01374	0.14101	1.31 X

```
378  40.0  0.41239  0.42455  0.01469  -0.01216     -0.11 X
379  39.0  0.27163  0.41539  0.01374  -0.14375     -1.34 X
380  35.0  0.64417  0.37873  0.01011   0.26544      2.46R
385  35.0  0.62023  0.37873  0.01011   0.24150      2.24R
389  40.0  0.31692  0.42455  0.01469  -0.10763     -1.00 X
390  39.0  0.35307  0.41539  0.01374  -0.06232     -0.58 X
391  39.0  0.32927  0.41539  0.01374  -0.08611     -0.80 X
392  40.0  0.43858  0.42455  0.01469   0.01403      0.13 X
398  40.0  0.34337  0.42455  0.01469  -0.08117     -0.76 X
400  39.0  0.37976  0.41539  0.01374  -0.03562     -0.33 X

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.
```

Explain the nature of this list in this case (consider the length of the list and the values in it.)

d) **(2 marks)** The company wants to estimate the proportion of 40 year old copper pipes with minimum wall thickness less than 0.15mm, because such pipes are considered to be at elevated risk of leaking. Produce such an estimate if you can, of if you cannot, explain why not.

2. **(5 marks total)** The natural gas company from the previous question has additional information about the 400 pipe samples. For each sample it also has x_2, the average rate of flow of gas that went through the pipe, and x_3, the acidity level of the soil in which the pipe was buried.

The company believes there might be an interaction between age x_1 and flow x_2, so it will consider fitting the model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{1i} x_{2i} + \varepsilon_i$$

to the data resulting in an R^2 value of 0.241 (which Minitab would call "24.1%".)

a) **(2 marks)** Produce the Analysis of Variance table for this multiple linear regression model fit, filling in the blanks in the following table:

Analysis of Variance

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

b) **(3 marks)** The second and third stages of a "forward" sequential model fitting strategy resulted in the following computer output (with the "Constant" lines of the tables omitted for clarity):

```
Second stage:

Predictor      Coef   SE Coef     T      P
x1         0.009160  0.001025  8.93  0.000
x3          0.00462   0.00177  2.61  0.009

Third stage:

Predictor      Coef   SE Coef     T      P
x1         0.011160  0.005957  1.87  0.061
x3          0.00389   0.00155  2.51  0.012
x2          0.00762   0.00492  1.55  0.121
```

Briefly describe the likely possible relationships that may exist among these three input variables and the output variable y.

3. **(10 marks total)** A manufacturing process is supposed to produce bolts of length 5 cm. If the mean length changes enough in either direction, the bolts will not be usable. Experience shows that the the bolt length has a Normal distribution with standard deviation $\sigma = 0.07$cm. The company plans to take a sample of size $n = 20$ bolts to see if they are being produced to the correct length, by testing the following hypotheses at the $\alpha = 0.01$ level:

$$H_0 : \mu = 5 \qquad H_a : \mu \neq 5$$

a) **(3 marks)** Derive the rejection region for this hypothesis test.

b) **(4 marks)** Suppose that unbeknownst to the company, the mean bolt length had actually shifted to 5.1cm when they took their sample. What is the probability that their hypothesis test would reject the null hypothesis under these circumstances? (In other words: what is the *power* of the test?)

c) **(3 marks)** What sample size would have been required so that the hypothesis test could have detected a difference of 0.1cm with power 0.9?

4. **(10 marks total)** A government scientist is going to compare four drugs that treat high blood pressure. Call the drugs A, B, C, and D. Each drug will be given to 20 patients and the reduction in blood pressure will be observed.

In advance of running the experiment the scientist is interested in comparing A with the average of B, C, and D. He is also interested in comparing whichever two drugs happen to have the largest average responses.

He runs the experiment and observes a total sum of squares of 707.45 and an error sum of squares of 553.46. He observes the following drug sample averages and sample standard deviations:

Level	N	Mean	StDev
A	20	3.472	1.995
B	20	6.280	2.665
C	20	4.525	3.160
D	20	6.972	2.840

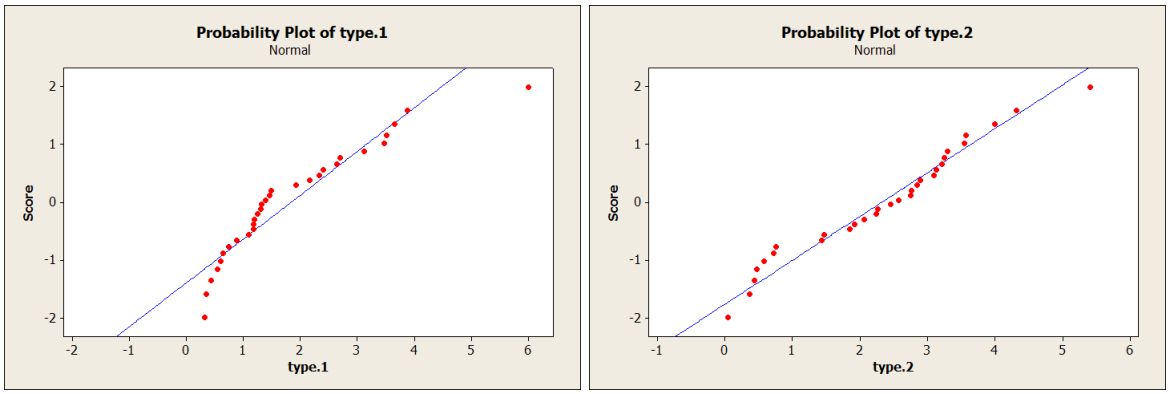
You have four tasks. Perform the hypothesis test to answer the general question "is there any difference among the drugs", and perform the two comparisons that the scientist had also wanted to perform as described above. Finally, assuming that a normal quantile plot of the residuals is nearly a straight line, comment on the extent to which the model assumptions have been satisfied.

(. . . question 4 space continued.)

5. (5 marks total) A company suspects that the configuration of the rail cars used to ship a product may have an effect on the moisture level of the product when it reaches its destination. The company collects samples of product for each of the two types of rail car (Type 1 and Type 2). Both samples contain 30 observations.

The Type 1 data has a sample mean of 1.85 and a sample standard deviation of 1.32. The Type 2 data has a sample mean of 2.05 and a sample standard deviation of 1.33.

Here are normal quantile plots of the two samples:

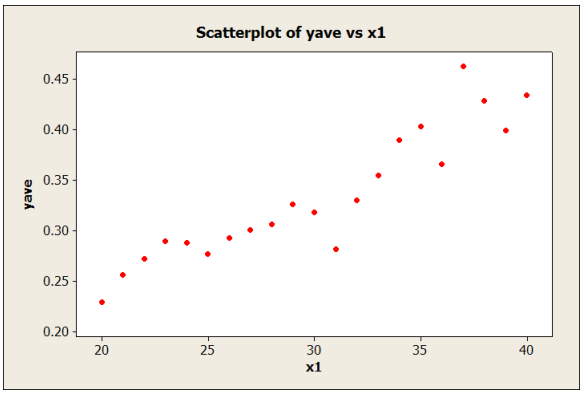


Test the hypotheses $H_0 : \mu_1 = \mu_2$ versus $H_a : \mu_1 \neq \mu_2$ by computing a p-value. Comment on the extent to which the model assumptions have been satisfied and whether or not any violations would affect the validity of your conclusion.

6. **(5 marks total)** An experienced instructor notices that 25% of students in a large undergraduate statistics course get a final mark of 80 or above whenever he teaches it. At the beginning of one such course, a sample of 50 students is selected from those about to take the course, and 10 of them predicted that they would get a final mark of 80 or above. Assuming that 25% of students really will get 80 or above as usual, is there evidence that these students did not accurately predict their performance in this course? Use a p-value in your conclusion.

7. **(5 marks total)** Here are two unrelated questions relating to regression analysis that you may not wish to attempt until you have completed the rest of the exam to your satisfaction.

a) **(2 marks)** Consider again the scenario in question 1 involving the gas company's copper pipes. Recall that pipes with minimum wall thickness are considered to be at high risk of leaking. Here is a new plot of the data, where this time the response variable is y^{ave}, the average minimum wall thickness for all pipes that are the same age.



Briefly explain why this is not a good way to analyze the data if your main purpose of the study is to assess the relationship between age and the number of pipes at high risk of leaking.

. . . this is probably more than enough space to answer this question!

b) **(3 marks)** Suppose you have a set of bivariate numerical measurements $\{(x_1, y_1), \ldots, (x_n, y_n)\}$.

A linear transformation $w = ax + b$ is applied to the x-coordinate of the data. Denote this transformed dataset by $\{(w_1, y_1), \ldots, (w_n, y_n)\}$, where $w_i = ax_i + b$.

Show that $r_{xy} = r_{wy}$, where r denotes the sample correlation coefficient. This shows that r is "unitless".