1. A dataset with sample size 40 consists of a response variable y and four (possible) input variables x_1, x_2, x_3, x_4. So 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	0.1554	-	-	-	9.3
	2	-	0.0075	-	-	17.8
	3	-	-	0.0009	-	25.6
	4	-	-	-	0.0001	31.9
2	5	0.0635	0.0090	-	-	24.8
	6	0.1859	-	0.0028	-	29.0
	7	0.5719	-	-	0.0010	32.4
	8	-	0.0210	0.0025	-	35.7
	9	-	0.0000	-	0.0000	62.5
	10	-	-	0.1964	0.0269	34.9
3	11	0.1830	0.0220	0.0068	-	38.8
	12	0.7656	0.0000	-	0.0000	62.6
	13	0.5286	-	0.1910	0.0629	35.6
	14	-	0.0000	0.2887	0.0000	63.7
4	15	0.6332	0.0000	0.2664	0.0000	63.9

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 '-'.

(a) Which of the two input variables are probably the most highly correlated and why?

(b) If you were to initially apply the forward regression strategy, which would be the first two models you would select, and why?

(c) Choose a final model using a general sequential method (possibly with forwards and backwards steps, if necessary) and briefly justify your choice.

(d) $\sum \left(Y_i - \overline{Y}\right)^2$ is 75.36. For the model you chose in (c), perform the hypothesis test using an F distribution that answers the question "is there any significant linear relationship between the inputs and the response".

2. 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 (recorded as variable x_1) 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 the Minitab output with some numbers removed:

```
Predictor      Coef   SE Coef     T      P
Constant    0.65801   0.02785 23.63  0.000
x1          *******   ******* *****  *****


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


Analysis of Variance

Source           DF      SS       MS      F      P
Regression      ***  1.3139  *******  *****  *****
Residual Error  ***  ******  *******
Total           ***  5.9843
```

(On the 2012 exam this was the first question. The second question was as follows. . .)

. . . 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.

(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) 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. A dataset with sample size $n = 63$ consists of an output variable y and five (possible) input variables x_1, x_2, x_3, x_4, x_5. There are 32 possible linear regression models of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$$

with $\varepsilon \sim N(0, \sigma^2)$. Excluding an x_i from a model is done by setting its β_i to 0.

Here are the sample correlation coefficients for all pairs of variables, both input and output:

```
          y       x1       x2       x3       x4
x1    0.156
x2    0.798    0.023
x3    0.292   -0.118    0.346
x4    0.794    0.251    0.596    0.143
x5    0.674   -0.143    0.888    0.249    0.579
```

(a) **(3 marks)** The final model that you would select using a general sequential method (possibly with forward and backward steps, if necessary) would definitely consist of one of three models summarized here:

Model	Output	R^2

1	`Predictor      Coef  SE Coef     T      P`
	`Constant   0.86870  0.09879  8.79  0.000`
	`x2         0.61177  0.08863  6.90  0.000`
	`x4         0.60958  0.09010  6.77  0.000`
		79.4%

2	`Predictor      Coef  SE Coef      T      P`
	`Constant    0.9252   0.1301   7.11  0.000`
	`x2          1.1474   0.2036   5.64  0.000`
	`x5         -0.1050   0.1066  -0.99  0.328`
		66.3%

3	`Predictor       Coef  SE Coef      T      P`
	`Constant    0.88132  0.09562   9.22  0.000`
	`x2           0.9028   0.1531   5.90  0.000`
	`x4          0.63633  0.08784   7.24  0.000`
	`x5         -0.18098  0.07892  -2.29  0.025`
		81.1%

Give a brief possible explanation of the behaviour of the p-value associated with β_5 in these three models.

(b) **(2 marks)** Which of the three models would you choose, and why?

(c) **(3 marks)** The total sum of squares is 176.3. For the model you chose in b), perform the hypothesis test using an F distribution that answers the question "is there any significant linear relationship between the inputs and the response".