Statistics for Engineering, 4C3/6C3

Take-home final exam: 2010

Kevin Dunn, dunnkg@mcmaster.ca **March 2010**

The purpose of this final exam is no different to other exams. The only difference is that this exam will also assess your ability to answer more realistic problems that require more time, thought, and access to a computer. The questions in this exam use actual data sets, which is one of the aims of this course. You are expected to use any appropriate tools to solve the problems in this exam, particularly the tools learned in this course. You may, however, use any software packages and tools to help solve the problems, as long as they are appropriate.

Some other points are:

- Complete this exam with 1, 2 or 3 other people. Groups greater than 4 will not be accepted; you don't need to do the exam in a group.
- 600-level students may complete this exam on their own, or with 1 other person.
- Please identify all your group members and sources of reference in your answer submission. **Please note: submit one paper per group**.
- The intention of the group work is that you discuss the questions and collaborate in the same way you have done with the assignments.
- Do not share any electronic files (e.g. Word documents, source code, Excel files) outside your group; take care in the computer labs to safeguard your work.
- Like any other exam, neither the TA nor myself are able to answer direct questions about the exam. Similarly you should not look for help about a specific question from other resources (e.g. asking for help with the question on a website, friends, etc).
- You may use the course notes and any other textbooks and resources though.
- There is no make-up for this exam.
- You will benefit from going through parts 5, 6 and 7 of the R tutorials on the course website.
- Your answers should preferably be typed up.
- Hand out date: 26 March 2010,
- **Hand in date: on or before 14:30 on 05 April 2010**
  - Paper hand-in: by **14:30 on 05 April, in class**
  - Electronic hand in: email to dunnkg@mcmaster.ca before the above time and date.
  - Electronic hand in must be in PDF format only (no Word, Excel or R files). If you are on Windows, you could use PDFCreator to make PDF's; Macs and Linux have built-in PDF capability.

Note: There are a maximum of 15 grade points available on this exam, plus a variable amount of bonus points in question 2. The 400-level students will be graded out of 12 points; 600-level students will be graded out of 14 points.

Question 1 [5]

The data set on the course website is from an actual designed experiment run at an oil company. The four factors under investigation were the levels of various additives in order to (i) learn more about the process and (ii) achieve a required volumetric heat capacity.

1. A full factorial model would allow you to estimate 2^4 terms (effects). Why are you unable to independently estimate all the effects? Comment on the actual levels at which the experiment was run at.
2. Calculate a predictive model that the company could use to estimate the response variable. Your model should only use main effects and 2 factor interaction terms.

   Hints:

   - You can use `as.numeric(C)*2-3` to convert column C to `-1` and `+1` values.
   - You can use the `ffFullMatrix` function with the `maxInt` input option to expand the given experimental design matrix. Or you can expand it manually in R or MATLAB.
   - Once you have your $\mathbf{X}$ matrix, you can use the `lm(...)` function to fit your linear model, like any other linear model, or you may use $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.
3. Refine your model from the previous part of this question to use only the significant terms. Use any remaining degrees of freedom to estimate confidence intervals for each of the significant terms.
4. The company wants to manufacture a fluid with a volumetric heat capacity as close to 33.2 as possible. Suggest a combination of process factors, in terms of **A**, **B**, **C** and **D**, that you predict would achieve this requirement. Give a rough estimate of the prediction error at your suggested settings. Show your calculations and your final answer should be in real-world values, not coded values. (This question could be made more interesting by assigning dollar values to each factor and finding the minimum cost specification; but we'll leave that fun for the 4G3 course).

Answer

Here is the data table for this question; the last four columns are calculated as explained below.

A	**B**	**C**	**D**	y	**A**	**B**	**C**	**D**
4.38	1.6	No	2.18	28.03	-1.00	-1.00	-1	-1.00
6.2	2.26	Yes	2.18	37.25	1.00	1.00	1	-1.00
6.2	2.26	Yes	2.18	37.22	1.00	1.00	1	-1.00
4.38	2.26	Yes	3.08	37.06	-1.00	1.00	1	1.00
6.2	1.6	Yes	3.08	39.23	1.00	-1.00	1	1.00
6.2	1.6	No	3.08	30.86	1.00	-1.00	-1	1.00
6.2	1.6	No	3.08	30.89	1.00	-1.00	-1	1.00
6.2	1.6	Yes	3.08	39.21	1.00	-1.00	1	1.00
4.38	2.26	No	3.08	29.31	-1.00	1.00	-1	1.00
4.38	2.26	Yes	3.08	37.03	-1.00	1.00	1	1.00
4.38	2.26	No	3.08	29.24	-1.00	1.00	-1	1.00
6.2	2.26	No	3.08	34.63	1.00	1.00	-1	1.00
5.60	2.26	No	2.78	29.73	0.34	1.00	-1	0.33
6.2	1.82	Yes	2.48	37.89	1.00	-0.33	1	-0.33
6.2	2.04	No	2.18	31.71	1.00	0.33	-1	-1.00
5.60	1.6	No	2.18	29.00	0.34	-1.00	-1	-1.00
4.98	1.6	Yes	2.18	36.15	-0.34	-1.00	1	-1.00
5.60	2.26	No	2.78	29.77	0.34	1.00	-1	0.33
6.2	2.04	No	2.18	29.40	1.00	0.33	-1	-1.00

1. The first part of this question requires you to convert the data from the actual values into the usual -1 and +1 values. You may have done this in MATLAB, or some other program, but in R you can do it with this code. The results are shown as the last four columns in the above table.

   CODE HERE

   You should notice that the experiment was not run in an orthogonal manner. We cannot independently estimate all the 2^4 effects because, even though we have more data points (19) than parameters to estimate (16), the experiments were not run in an orthogonal manner. In other words the $\mathbf{X}'\mathbf{X}$ matrix will not be a diagonal matrix, so the parameters cannot be estimate independently. There are also duplicate rows, so the true number of independent rows is less than 19.

2. A predictive least squares model can be calculated from the 19 experimental data points, even though the experiment was not done in a factorial manner. This is important to remember in your career as an engineer: many times the full set of experiments will not be completed (you run out of money, the objectives will change before you can complete them, you run into constraints *etc.*). But you can always calculate a linear model; it may not have all the great properties of a model from a full factorial, but it likely will be a useful model (see part 3 of this question).

   So to predict the response variable using the 4 main effects and the 6 two-factor interactions, we write this model

   $$y = b_0 + b_\text{A}x_\text{A} + b_\text{B}x_\text{B} + b_\text{C}x_\text{C} + b_\text{D}x_\text{D} + b_\text{AB}x_\text{AB} + b_\text{AC}x_\text{AC} + b_\text{AD}x_\text{AD} + b_\text{BC}x_\text{BC} + b_\text{BD}x_\text{BD} + b_\text{CD}x_\text{CD} + e$$

   The next step is to set up the $\mathbf{X}$ matrix and $\mathbf{y}$ vector in order to solve for $\mathbf{b} = \mathbf{X}'\mathbf{X}\mathbf{X}'\mathbf{y}$.

   $$\mathbf{b}' = \begin{bmatrix} b_0 & b_\text{A} & b_\text{B} & b_\text{C} & b_\text{D} & b_\text{AB} & b_\text{AC} & b_\text{AD} & b_\text{BC} & b_\text{BD} & b_\text{CD} \end{bmatrix}$$

   One of the first things to do is convert the experimental data into -1 and +1 factors. Duplicate points can be averaged ahead of time to get fewer rows. While you will get the same parameter estimates, this is not advised, because you loose degrees of freedom. The repeated experiments were true repeats; and so they provide extra degrees of freedom with which to estimate the confidence intervals.

   So to set up the corresponding $\mathbf{X}$ and $\mathbf{y}$ in R, one could write (there are other ways also):

   we have to set up the matrices to calculate the parameters using least squares.

3. This question was actually more of a least squares question than a DOE question.

   that the company could use to estimate the response variable. Your model should only use main effects and 2 factor interaction terms.

   Hints:

   - You can use `as.numeric(C)*2-3` to convert column C to `-1` and `+1` values.
   - You can use the `ffFullMatrix` function with the `maxInt` input option to expand the given experimental design matrix. Or you can expand it manually in R or MATLAB.
   - Once you have your $\mathbf{X}$ matrix, you can use the `lm(...)` function to fit your linear model, like any other linear model, or you may use $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

#. .. rubric:: Question 2 [6]

This question considers a pilot-plant for a pulp-and-paper process where you are investigating three factors to optimize a new product:

- **A** = cooking temperature
- **B** = amount of sodium hydroxide added
- **C** = presence or absence of a certain cellulose

The response variable of interest is a quality variable, which is then converted to a profit value, based on current market conditions for this product. There are some constraints on the system:

- **A**: cooking temperature can only be between 390 to 480K
- **B**: between 5 to 30 kg of sodium hydroxide can be added
- **C**: can either be absent or present
- $\mathbf{A} + 2\mathbf{B} \leq 500$ is a safety constraint that avoids high temperatures and high quantities of sodium hydroxide.

Your manager has allocated a budget for 20 experiments ($10,000 per experiment), and your objective is to find the most profitable operating point with the fewest number of experiments. The current baseline operation is at **A** = 440K , **B** = 20 kg of NaOH and **C** = Absent. These 3 factors can be moved anywhere, as long as they obey the constraints.

You are expected to use all the tools learned in this course to solve this problem; in particular: using clear visualization plots, linear models, tests of significance, design of experiments and response surface methods.

A model of the process has been computerized, and is available on the course website. Nominate *one of your group members* and email their name and student number to the course instructor, together with the names of the other group members. This will activate an account for your group. Once you sign into the account you will be able to specify the levels of the 3 factors and the server will return the response (i.e. the server will "run" the experiment for you).

The 6 grading points for this question will be marked only on the systematic methodology used to approach the optimum. Make sure you explain your approach clearly, exactly as you would have to in a report to your manager. Your final answer must report (i) the operating levels at the optimum for factors **A**, **B**, and **C** as well as (ii) the expected profit at this optimum.

There are bonus points available, depending on the number of runs you *don't use* and your closeness to the optimum, according to this formula:

Extra grades = $1.5 \times \dfrac{\text{Your optimum - Baseline}}{\text{True optimum - Baseline}} - 0.15N + 3.0$, where N is your number of experiments.

Please note:

- there is error (noise) in the response variable;
- the server will return different results for different groups;
- please enter your conditions carefully - if you use the wrong settings you will have to work with those results;
- the server will also keep track of and display all your previous experiments on a results sheet.

Once you have completed the question, print out the result sheet from the server and submit that with your answer. (The true optimum and operating point for the optimum will be be available after the midterm is handed in).

Answer

Question 3 [4]

This question is based on an actual problem experienced at a company, but the data are simulated, and simplified only very slightly, to reflect the essence of the problem.

Recent trends show that the yield of your company's flagship product is declining. You are uncertain if the supplier of a key raw material is to blame, or if it is due to a change in your process conditions. You begin by investigating the raw material supplier.

Data on the course website provide the 6 values that are known for each lot of raw materials: 3 of them are a size measurement on the plastic pellets, while the other 3 are the outputs from thermogravimetric analysis (TGA), differential scanning calorimetry (DSC) and thermomechanical analysis (TMA), measured in your laboratory. These 6 measurements are thought to adequately characterize the raw material. Also provided is a designation "Adequate" or "Poor" that reflects the process engineer's opinion of the yield from that lot of materials.

1. How would you go about finding out, univariately, which variable(s) are associated with the declining yield. Provide an illustration (you may choose to use the actual data, but this is not required).
2. Build a latent variable model from these data. Colour-code the plots (or use different marker styles) with the designation of "Adequate" or "Poor". Are there any features of interest in the plot?
3. What would be your recommendation to your manager to improve the yield?
4. Could you save your your laboratory some analysis time and money by eliminating any redundant measurements on the raw materials?

Answer

1. There are several alternatives here. One of the most effective ways is to plot two box plots, one for the each variable; separate box plots for the "Adequate" and "Poor" samples. Other suggestions could be to perform a t-test for the between-group difference and calculate the z-value (risk). Many people also used a scatterplot matrix of the raw data to effectively identify the univariate relationships of the six variables with "Outcome".

   It wasn't required, but most people showed that high size values, and small TGA and TMA values had the most relationship with poor outcome. The DSC value did not seem to impact the Outcome.

2. A latent variable is constructed using the 6 variables in the PCA model. The first three components accounted for 44.5%, 34.0% and 14.0% of the variance, for a total of 92.4% of the variance explained with 3 components. Even this third component may not be too important to solving our problem.

   Interpretation of of the 1st component is mainly due to the 3 size variables, which are strongly correlated. The second component mainly explains the TMA and TGA variables (thermal effects), while the 3rd component is mainly for the DSC effect.

   The score plots are shown here for t_1 against t_2, with different colours and shapes for the Poor and Adequate designation. Its clear then that low values of t_1 and higher values of t_2 are associated with poor yields.

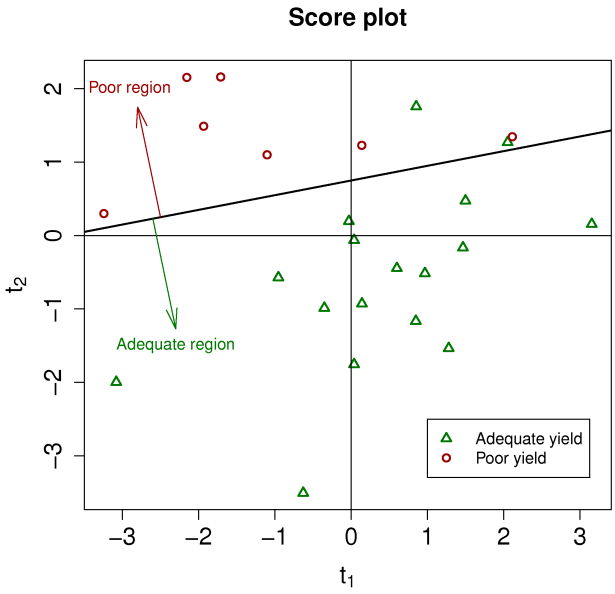
3. Almost all adequate yields were at high values of t_1 and low values of t_2. But we must "translate" those scores back to real-world quantities that our colleagues and managers can work with. That means we should run our process with smaller sized plastic pellets (we get that from t_1) and the properties of those pellets should be such that we get higher values from the TGA and TMA. The DSC value doesn't have too much of an effect.

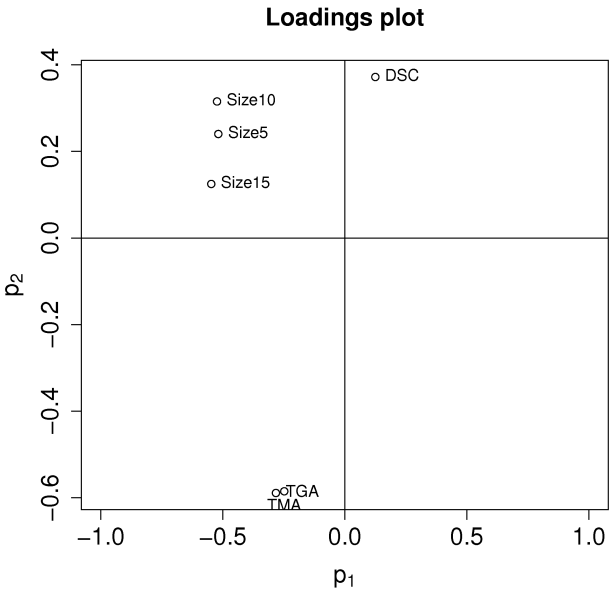
   One can construct regions in the score plot as shown in the above figure. If a batch of material is received that is in the Poor region, then it could be used in another part of the process that is not too sensitive to this problem. Or the batch could be returned to the supplier, or some other arrangement made, such as blending two batches to achieve a better set of raw material properties.

4. As the loadings plot shows, the TMA and TGA values are very highly correlated. One of these two measurements could be omitted, based on economics; perhaps one of the instruments gives additional information (e.g. measures impurities also), or one has a lower measurement error than the other.

   One probably has to retain all the size measurements; they are usually measured on the same device at the same time. But it they were separate measurements, then you could omit one of the three.

   Finally, I would still retain the DSC value; while it is not so informative to this particular problem, it is an important 3rd component. Perhaps there is some other quality aspect that DSC measured that isn't related to the Poor vs Adequate designation.


Score plot
Poor region
Adequate region
Adequate yield
Poor yield
t1
t2


Loadings plot
DSC
Size10
Size5
Size15
TGA
TMA
p1
p2