Statistics for Engineers, 4C3/6C3

Take-home final exam and DOE project

Kevin Dunn, dunnkg@mcmaster.ca **Due date: 28 March 2011**

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 realistic problems that require more time and thought, will require you to read up on other, similar statistical tools learned in the course, and require using computer software. Much in the same way you will solve day-to-day problems as an engineer when you graduate soon.

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.

Important notes

- Complete this exam with 1 or 2 other people. Groups greater than 3 will not be accepted.
- You don't need to do the exam in a group, please work on your own, if you prefer.
- 600-level students may complete this exam on their own, or with 1 other person.
- Identify all your group members *and sources of reference* in your answer submission. **Note: one paper/printed submission 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 though.
- There is no make-up, nor extended time granted for this exam.
- Your answers should preferably be typed up.
- Hand out date: 20 March 2010; hand in date: on, or before, **10:30 on 28 March 2010**. Neither electronic hand-ins, nor late hand-ins will be accepted, unless I have given your prior written permission.

Grading

The 400-level students will be graded out of 100 points; 600-level students will be graded out of 105 points, and 600-level students are expected to show insight and technical accuracy at the graduate student level. This take-home final exam and DOE project count 25% of the overall course grade.

Question 1 [10 + 5 (for 600-level students)]

Sequences of data record the level (height) of a particular raw material in a tall storage container at your company. The level recording is both inaccurate and infrequent, since it is manual. The measuring device is recalibrated from time-to-time. The container's height is about 30 meters.

The container is refilled with plastic pellets, by railcar 5 days after your company places an order with the material supplier. The problem here is a logistics problem, also called a supply-chain problem: if we order too soon, then there isn't enough space to store the material delivered; if we order too late, then we risk running out of the raw material before the next delivery.

You have been asked by your boss to build a model that predicts the *rate of material use* in the container. That way we can make a prediction of the storage level in the future, and plan ahead to place an order for more material, just in time.

Data on the website give you the height measurements, in meters, for January, February and March. Use these data to develop a prediction strategy that will answer this question: *what is your prediction of the container level on 26 March 2011* (i.e. 5 days from the last available height measurement) if its not refilled during the week of 21-25 March?

Notes:

- This problem is similar to one that your instructor recently worked on with a client, and the data are nearly the same as from that problem.
- To solve this problem we suggest you plot the data, and decide on which tools to use from this course.
- *Hint*: consider using a device like the `datenum` function in MATLAB, although you don't have to use MATLAB for this question.
- Grading for this question is purely on methodology, so you should outline your strategy clearly, and the assumptions used to solve the problem.
- 600-level students: I expect you to research and use tools like robust least squares, which are similar to ordinary least squares, but were not taught in this course. In your answer you should contrast the ordinary least squares modelling to robust least squares modelling.

Question 2 [15]

This dataset contains the grades for a class of students at McMaster University, in Chemical Engineering. The recorded values are the average of sub-components: e.g the `Tutorial` variable is the average of all tutorials.

Also available in the data, as the `Prefix` column, are the first 2 digits of the student number, which are assigned in the year the student first enrolls at McMaster. For example, a `06` indicates the student started in 2006. To a rough approximation this reflects the student's age (maturity).

This particular course permitted students to work in groups for assignments, tutorials and the take-home exam. The groups were self-selected, and varied during the semester.

Of interest is whether the assignments, tutorials, midterms or take-home exam are a good predictor of the student's performance in the final exam. Use these data to answer the questions:

1. How would you characterize the distribution of the final exam grades?
2. Which variable in the data set is the strongest predictor of the student's final grade? Describe how you come to this conclusion.
3. Is there any evidence to support the hypothesis that the student's maturity has an effect on their final exam grade? Consider only the students enrolled in 2007 vs 2008.
4. After a multiple linear regression of the final exam grade on all the other factors, show one or two plots to diagnose the residuals. Are there any problematic observations that you would investigate and perhaps eliminate from the model?
5. Of interest are those students that make a large improvement from their `Midterm` exam to the `Final` exam grade. Calculate this difference variable and regress it on the remaining variables in the data set to verify if there is some indicator that will identify which students are expected to improve.

Question 3 [18]

A vineyard grows several varieties of grapes for various wines they produce. They use factorial experiments in all their operations: to investigate conditions that generate the highest level of profit in their wine store (lighting brightness, natural light, shelf height and locations, free vs paid tastings, *etc*), and their on-site restaurant (amount of hourly wage paid to servers, colour of decor, lighting levels, *etc*). On the farm lands they continually investigate factors that minimize equipment maintenance (brand of motor oil, frequency of preventative maintenance, quality of diesel fuel used, *etc*).

This current season they used a fractional factorial to develop a wine. The cost of doing these experiments, compared to the experiments mentioned above, is high. So they had to get them right, and run as few as possible. Appropriately, they have used a fractional factorial in these factors:

Factor	Factor investigated	Low level	High level
A	Grape clone	Edeltraube	Traminec
B	Oak type	Acacia	Limousin
C	Age of barrel used	Old	New
D	Yeast brand used	Epernay	Pasteur
E	Stems included	None	All
F	Level of barrel toasting *	Light	Medium
G	Whole cluster pressing of grapes	None	10% of total
H	Fermentation temperature	24°C	33°C

$*$ When a wine barrel is constructed, the cooper (barrel maker) will place the partially assembled barrel over a fire and "toast" the barrel. Winemakers can request barrels with light, medium and heavy toast, which has an effect of the wine's taste.

The results of the experimental trials are given on the website where the first 8 columns represent the levels of the 8 factors, and the last 5 columns are the taste values from 5 judges. The experiments and tastings were performed in random order.

1. Could the winemakers have run fewer than 16 experiments? Please explain.
2. What is the resolution of this design?
3. What does this resolution imply about the results we will get in the subsequent analysis?
4. Use an average of the tasting panel as the response variable and calculate which of the 8 factors have the greatest effect on the taste (higher numbers are better). Also report which two factor interactions are aliased with these important main effects.
5. Calculate the standard deviation of the 5 taste values from each judge and determine which conditions lead to the most robust taste response (wine that produces a consistent response from all judges).
6. Is there a combination of factors in the experiments that leads to both high taste scores and good consistency among all judges?
7. Which factors would you suggest the winemaker omit from next year's experimental trials and hopefully reduce the cost of their experiments?
8. The winemakers know the wine produced from the experiment with all levels at their high value is unlikely to be a good wine, and this was confirmed in the experiment. How could you obtain the same fractional factorial (8 factors in 16 runs, with the same resolution), that does not contain this experiment?

Question 4 [20]

This question considers a batch reactor process where you convert the raw materials to a low-volume, but very valuable product. There are 3 major factors that can be adjusted:

- **A** = temperature during the batch reaction, Kelvin
- **B** = batch duration in minutes
- **C** = choice of solvent: acetone or xylene

The response variable is the percentage conversion calculated by weighing the final product vs the raw material weight. This conversion is directly related to the process profit, so its the only variable that you need to optimize for by adjusting the 3 factors.

There are some constraints on the system:

- **A**: the batch temperature must lie between 390 to 480K.
- **B**: the batch duration must be longer than 20 minutes, but less than 50 minutes before unwanted side reactions start to consume the product.
- **C**: either acetone or xylene can be used
- $4\mathbf{A} + 9\mathbf{B} \leq 2250$ is a safety constraint that avoids simultaneous high temperatures and long reaction durations.

The first baseline experiment is at **A** = 463K , **B** = 28 minutes and **C** = Acetone. The 3 factors can be moved anywhere, as long as they obey the constraints. You have a budget of 20 experiments (about $8,000 per experiment), and your objective is to find the operating point that gives the highest conversion.

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

A simulation 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 grading for this question will be marked mostly on the **systematic methodology** used to approach the optimum, exactly as you would have to in a report to your manager.

Since the cost of each experiment is so high, you must explain your approach clearly, justifying why you pick each experiment, and what you plan to do with each new experiment's result. In particular, you should use your models to predict the result of the next experiment before you run it (of course this doesn't apply to the first few experiments).

Your final answer must report (a) why you decided to stop with the particular number of experiments you actually ran, (b) the optimum operating levels for factors **A**, **B**, and **C** that you will recommend to your manager and (c) the expected conversion 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 = $3.0 \times \dfrac{\text{Your optimum} - \text{Baseline}}{\text{True optimum} - \text{Baseline}} - 0.25N + 5.0$, where N is your number of experiments.

Please note:

- There is calculation error (noise) in the response variable in the order of ± 0.75 percentage points.
- For the same levels of **A**, **B** and **C**, the simulation 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.
- You must wait 1.5 hours between each experimental condition; please plan your time accordingly.

- 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 and submit that with your answer. (The true optimum and operating point for the optimum will be be available after the exam is handed in).

DOE project [37]

As described in more detail in the project handout, the grading will be for:

- Describe your objective for the system under investigation. What is/are the outcome variable/s you are investigating; how are they measured?
- Outline the factors that you expect will influence the outcome variable. How will you measure the factors, over what range will you vary them? State how you expect each factor to affect the response(s); do you expect any interactions?
- Disturbances: which factors are known to affect the response but not being investigated here? How do you control for them?
- Plan an experimental program that will change the system's factors and control for disturbances. Be specific on how you chose your design.
- Execute the experimental program, logging all relevant details (e.g. experiments that are "weird", unusual events). Take photos/keep a log sheet.
- Analyze the experimental results using the tools introduced in the course.
- The conclusions, related back to your original objectives. What would be the next set of experiments you run?

The project is handed in separately, either on 28 March or 31 March, at your preference. Please submit a printed hand-in, *but also an electronic copy of your DOE report* (PDF preferred) to me.

END