Statistics G001 In-Course Assessment, 15 November 2016

- Your solutions should be your *own work* and are to be handed in to the departmental office before (but not after) TUESDAY, 22 NOVEMBER 2016, 4 pm. Unless there are extenuating circumstances, ICAs submitted after the deadline will be penalised. For instance, for submissions up to 2 working days late a deduction of 10 percentage point will be applied. For more than 2 but less than 5 working days late, the ICA will be capped at a pass. For more than 5 working days late the mark will be 0 but the ICA will be recorded as completed.
- Before you hand in your work, complete and sign the slip below this rubric, cut it off and attach it firmly to your work. The slip has to be signed by all group members. (See the next page for the rules on group submissions.)
- Please make sure that it is recorded on a list of students of this course that you have handed in your work.
- Non-submission of in-course assessment may mean that your overall examination mark is recorded as "non-complete", i.e. you might not obtain a pass for the course.
- Any plagiarism will normally result in zero marks for all students involved, and may also mean that your overall examination mark is recorded as non-complete. Guidelines as to what constitutes plagiarism may be found in the Departmental Student Handbooks.
- Your work will be returned to you for feedback and you will receive a *provisional* grade – *grades are provisional until confirmed by the Statistics Examiners' Meeting in October 2017.*
- You should return the marked work to the lecturer. Your work may be required for perusal by the visiting examiner.

I am aware of the UCL Statistical Science Department's regulations on plagiarism for assessed coursework. I have read the guidelines in the student handbook and understand what constitutes plagiarism.

We hereby affirm that the work we are submitting for this in-course assessment has entirely been carried out by us. Additionally, every parts of the solution is individually signed by the group member who wrote it down.

Signed:

Signed:

Signed:

Please print your name:

Please print your name:

Please print your name:

Date:

Note that **group submissions** are allowed for this ICA. The rules for this are as follows.

1. A single set of solutions can be handed in by groups of **up to 3 students**. This means that you can also submit your solutions either alone or as a pair or a group of three students.

2. **Students alone are responsible for forming groups.** The lecturer will neither decide about who works together, nor help with the organisation of such groups.

3. **All members of a group will get the same marks**, i.e., all will get marks for the whole set of solutions, provided that the conditions below are fulfilled. However, if rule 7 is violated, members of the group who didn't write down enough may not get the overall marks of the whole group.

4. Working together and discussing solutions *within groups* is fine and the usual plagiarism regulations don't apply to this. However, they do apply to plagiarism of work of other groups or other sources.

5. All group members must be indicated on all pages of the submission.

6. Group solutions should be handed in as a single, stapled document.

7. Additionally, **every part of the solution must be signed by the group member who wrote it** and **every group member must write down at least one block of the solutions alone.** (As long as this condition is fulfilled, other blocks can be split up between group members.)

   Explanation: your solution should comprise the following four "blocks" (you may need to read the questions to understand this):

   (a) Question 1, fitted models that you do not choose as your "recommended fitted model" and discussion.

   (b) Question 1, your "recommended fitted model" and discussion (discussion of model-independent plots that have some implications for the choice of the model or the fitting method may be included in block 1 and/or 2).

   (c) Question 1, report for the traffic institute.

   (d) Question 2.

1. The dataset `emissionssw.dat` can be found on the Moodle course page[1].
Load with `read.table("emisionssw.dat",header=TRUE)`.

The dataset gives 2022 measurements of NOx (nitrogen oxide) pollution content in the ambient air and some related variables. The measurements were taken over one year (typically 5-6 measurements a day) at a certain place in Switzerland close to a motorway. Data are sorted in order of the day and time at which the measurements were taken.

The variables are (in order of appearance as columns in the data file):

nox NOx concentration in ambient air [ppb].

noxem Sum of NOx emission of cars on this motorway (units not given in my source).

ws Wind speed in m/s.

humidity Absolute humidity in the air in g/kg air.

The data were collected by an environmental research institute in order to make quantitative statements about the strength of the influence of the three variables noxem, ws and humidity on the response variable nox.

You are asked to produce at least two fitted models and to come up with a single "recommended fitted model" which you think is the best for these data out of your fitted models.

You are asked to

- provide all relevant computer output including the R-code that produced it,
- write comments on the fitted models that you don't recommend, stating why you don't recommend them and what else you have learnt from them that was relevant for the design or interpretation of your recommended model, making clear reference to R-output and graphics,
- write comments on your recommended fitted model including diagnostic plots, making clear reference to R-output and graphics,
- write a report for the institute, explaining the relevant information (including reasons for doubt about its reliability if necessary) in a clear and understandable way.

2. Refer to the expression for D_i in page 26 of the lecture notes. Show that the influence vector $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}$ can be written efficiently as

$$\mathbf{C}^{-1}\mathbf{x}_i \frac{e_i}{1-h_{ii}},$$

where $\mathbf{C}^{-1} = (\mathbf{X}^T\mathbf{X})^{-1}$, $e_i = y_i - \hat{y}_i$, $\hat{y}_i = \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$, and h_{ii} represents the i^{th} diagonal element of the hat or projection matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ (as defined in page 23 of the lecture notes).

[Hint: Define $\boldsymbol{C}_{(i)}^{-1} = (\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)})^{-1}$, where $\boldsymbol{X}_{(i)}$ is the sub-matrix formed by omitting the i^{th} row of $\boldsymbol{X}$. The solution to $\boldsymbol{C}_{(i)}^{-1}$ can then be written as

$$\boldsymbol{C}_{(i)}^{-1} = \boldsymbol{C}^{-1} + \frac{\boldsymbol{C}^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{C}^{-1}}{1-\boldsymbol{x}_i^T\boldsymbol{C}^{-1}\boldsymbol{x}_i}.$$

Also, note that $\boldsymbol{d}_{(i)} = \boldsymbol{d} - \boldsymbol{x}_i y_i$ where $\boldsymbol{d} = \boldsymbol{X}^T\boldsymbol{Y}$. You might find it convenient to write down the expression for $\hat{\boldsymbol{\beta}}_{(i)}$ and then make all necessary substitutions and simplifications.]

[1] In case of difficulties with the access to the dataset, please contact the Lecturer, giampiero.marra@ucl.ac.uk

More information for question 1:

- By a **"fitted model"** it is meant that you fit a statistical model by one of the methods introduced in the course. The statistical models that you fit should be stated formally with definition of notation. You may fit the same statistical model (i.e., involving the same variables) by different methods, which counts as different fitted models.
- While you may fit as many models as you want "privately", please submit information (R-output and comments) for **at most THREE fitted models**. Full marks can be achieved by fewer than four fitted models.
- I do **not** expect you to comment on the parameter estimators and corresponding significant tests of the fitted models that you don't recommend. Generally the comments on these models can be brief and do not need to cover all aspects that could potentially be found, particularly not if they are not relevant to your attempts to improve the model or appear in the recommended model again and are discussed there.
- You are **not** expected to perform any variable selection techniques from Section 2.4 of the notes on these data.
- There is no unique optimal solution (I have a recommended fitted model myself but other fitted models may be equally good or better) and a good fitted model is not necessarily perfect in all aspects. Finding good models will be rewarded but note that it is more important (for marking) that the discussion of your fitted models, whatever their quality, makes sense. Particularly, you can still get high marks if you don't find a really convincing model but are honest and clear about the shortcomings of the one you recommend.

  Note that the dataset (as many real datasets) is somewhat "nasty" and it may well be that some clearly visible problems remain with any fitted model you could come up with.
- The discussion in blocks (a) and (b) (comments on non-recommended and recommended fitted models) may make use of statistical technical terminology, but the **report for the institute should not assume statistical knowledge** of the readers (mathematical knowledge on school level may be assumed; the report is for the institute, not for the general public). **The report for the institute should be self-contained. It should not make reference to any R-output**. If you want to include or make reference to graphs in the report, please include them clearly in the report (if you want to use a graph for the report that you have used before, it makes sense to submit it twice, namely together with your general R-output as well as in the report).
- The dataset is based on a real dataset that has been manipulated. Some information about it used in this ICA has been made up.

General:

- Do not write more than **SIX typed or NINE handwritten pages** (not including R-output and graphics) using a resonable letter size. Longer solutions will be penalised as well as irrelevant and unjustified statements (shorter solutions are fine).
- Handwriting, if used, has to be legible.
- Marking scheme: 60 marks for question 1, 40 marks for question 2. In question 1, all the blocks are intended to carry about equal weight, but marks will be assigned in a flexible way. For example, block (a) may carry a higher weight for those who fitted (and discarded) more models.

- Please note the warning on the cover page about the penalties for plagiarism. It is very easy to detect the copying of other group's written conclusions. Any detected copying will be severely penalised.