STAT0028 In-Course Assessment (ICA), 12th November 2024

The ICA is made up of two questions, which are given below.

1. The dataset `emissionssw.dat` can be found on the Moodle course page. You may want to load it 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 sequentially 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 [parts per billion, 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 regressors `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

(a) provide all relevant computer output including the R-code that produced it;

(b) 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;

(c) write comments on your recommended fitted model including diagnostic plots, making clear reference to R-output and graphics;

(d) given your final model, provided you have at least two regression coefficients excluding the intercept, how would you test the null hypothesis that the effects of two regression coefficients are the same? Comment on your empirical finding;

(e) write a report for the institute, explaining the relevant information (including possible reasons for doubt about its reliability) in a clear and understandable way.

2. Refer to the expression for D_i on page 27 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 on page 24 of the lecture notes).

[Hint: $\boldsymbol{C}_{(i)}^{-1}$ can be easily derived by recalling the definition of $\boldsymbol{C}^{-1}$. Alternatively, it can also be written as

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

Note also that $\boldsymbol{d}_{(i)} = \boldsymbol{d} - \boldsymbol{x}_i y_i$, where $\boldsymbol{d} = \boldsymbol{X}^T\boldsymbol{Y}$. You may find it convenient to start off by writing down the expression for $\hat{\boldsymbol{\beta}}_{(i)}$.]

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 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 at 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 a total of **SIX typed or NINE handwritten pages** for all three questions (not including R-output and graphics), using a reasonable 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, a block may carry a higher weight for those who fitted (and discarded) more models.