1 Instructions

You must write a report, in English. Work in groups of two. Discussion between groups is permitted (and encouraged), as long as your report reflects your own work. Write a clear report, presenting your approach to the assignment, discussing methods and results. Results discussion and interpretation is important. Just reporting results is not enough! It should be noted that for some questions there isn’t a unique “right” answer and there are a myriad of different issues that you could discuss, so use your imagination. In addition to the text, use as many figures and tables as necessary, with explanatory captions. The report should be readable, not a random disorganized collection of thoughts, plots and tables (see also the Peer Review Guidelines at the end of this document). For example, it should be possible for the reader to understand what you are doing without having access to your code. Also, key information may be better summarized in tables than by including the R printouts (e.g. it may be enough to give regression coefficients and p-values without all the accompanying information provided by R). There is no need to include your R code in the report, but you can include some of the R output.

Bring a printed version of your report to the peer assessment (April 24, 1315), or email the report to umberto@maths.lth.se at least 1 hr in advance so I can print a copy.

Form groups

Form groups of two students via “SAM”: https://matstat.sam.cs.lth.se/Labs

Final submission

E-mail the final version (a single PDF document) to one of the following addresses by the deadline 16.00 on 28 April. Also attach to the same message your R-files (or implementation in other language), with a file named proj1.R that can be used to run your analyses.

- FMSN30/MAEM22 students: email to FMSN30@matstat.lu.se
- FMSN40 students: email to FMSN40@matstat.lu.se

Subject field of the email: write “Project1 by studid1 and studid2” where studid1 and studid2 are the id numbers for two students in a given group (forgot your id? Go to the link in the “Form groups” section).

Example: Project1 by d08xhj and d08fjh

2 Maths education in Portugal

We are considering real-world data of students achievement in mathematics in secondary education. Data have been collected in two Portuguese public schools using mark reports and questionnaires in years 2005–2006. The sample size of respondent students is 395. The dataset studentmath.txt can be downloaded from the course webpage.
3 Questions

Several variables have been measured on 395 students, see section 4 for definitions. Our main goal is understanding which variables best “explain” the students final grade in mathematics (G3).

3.1 Basic descriptives

1. It is always a good idea to “get to know” your data by first looking at empirical distributions of your variables and possible associations between pairs of variables. Are there peculiarities in the material? If yes, describe them.

3.2 Simple linear regression

1. It should not be surprising that students grades in mathematics obtained during the year could help predicting the final grade in maths (G3). Investigate whether indeed there seem to be a linear relationship between $G_3$ and $G_1$ and between $G_3$ and $G_2$. Investigate whether requirements for linear regression seem to be fulfilled. In case of violations of some requirement try to see whether some transformation could help (notice such a transformation might not exist!).

2. Fit two models to explain the final grade, one using the first period grade as predictor and one using the second period grade. Discuss differences, if any. How do you interpret the value of the slope coefficient for each model?

3. What is the expected final grade for a generic student who received grade 13 in the first period? And how come a student receiving a grade equal to 1 in the first period is expected to receive a negative $G_3$ grade? How is this even possible? Explain.

4. Compute “by hand” (i.e. without using predict but you can still use R to perform some calculations) confidence bounds at a significance level $\alpha = 0.05$ for the expected final grade you computed in the previous question.

5. Compute “by hand” a suitable interval using $\alpha = 0.05$ for the final grade of an unobserved student assuming he/she gets a G1 grade of 13.

6. The interval computed in question 5 seems quite wide. Perhaps unsurprisingly, if we use the second period grade as covariate we obtain a reduced uncertainty in the resulting interval. Verify this and produce an explanation for this fact.

3.3 Multiple linear regression

1. Intuitively we might expect that the number of school absences to some extent do influence the final grade. You might have previously found that absences alone is not a relevant predictor. Use a statistical test to show that absences is useful when used in addition to G2. However, look at the surprising result that the coefficient of absences is suggesting. Why is that? This is non-trivial but try to investigate.

2. What is the estimated probability that a student who has previously failed exactly 1 class and in the current year has been absent 10 times and received a $G_2$ grade of 12 will get a final grade of at least 15? Hint: fit an appropriate model to estimate the final grade given the suggested covariates, then assume a Gaussian distribution centred at the estimated grade (this is just an approximation, but should work fine if the model is adequate).

3. It is very easy to passively rely on the R output to make conclusions, especially when only looking at p-values (this is what most do in reality, unfortunately). Suppose we want to explain $G_3$ given $G_2$, the number of absences, the number of failures and the travel time. Does travelling more than 1 hr seem to have an effect on the final grade, compared to having less than 15 min travel time? By how much? Does this imply
that the time spent travelling is an irrelevant/determining factor to explain G3 (given the other covariates already in the model)?

4. Use automatic identification tools to identify potentially relevant variables for explaining G3. Be critical and use automatic tools only to help you screen among many variables. Discuss your findings.

5. Conduct outliers detection and find (potentially) influential observations. Discuss your findings. For example give a motivation why an observations is an outlier and/or is influential.

4 Variables in the dataset

Here are the definitions for the variables considered in the dataset.

**school** - student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)

**sex** - student's sex (binary: "F" - female or "M" - male)

**age** - student's age (numeric: from 15 to 22)

**address** - student's home address type (binary: "U" - urban or "R" - rural)

**famsize** - family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)

**Pstatus** - parent's cohabitation status (binary: "T" - living together or "A" - apart)

**Medu** - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)

**Fedu** - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)

**Mjob** - mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "athome" or "other")

**Fjob** - father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "athome" or "other")

**reason** - reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")

**guardian** - student's guardian (nominal: "mother", "father" or "other")

**traveltime** - home to school travel time (numeric: 1 - < 15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - > 1 hour)

**studytime** - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - > 10 hours)

**failures** - number of past class failures (numeric: \( n \) with \( 0 \leq n \leq 3 \))

**schoolsup** - extra educational support (binary: yes or no)

**famsup** - family educational support (binary: yes or no)

**paid** - extra paid classes within the Math course (binary: yes or no)

**activities** - extra-curricular activities (binary: yes or no)

**nursery** - attended nursery school (binary: yes or no)
higher - wants to take higher education (binary: yes or no)
internet - Internet access at home (binary: yes or no)
romantic - with a romantic relationship (binary: yes or no)
famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
freetime - free time after school (numeric: from 1 - very low to 5 - very high)
goout - going out with friends (numeric: from 1 - very low to 5 - very high)
Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
health - current health status (numeric: from 1 - very bad to 5 - very good)
absences - number of school absences (numeric: from 0 to 93)

Here are the Maths grades for three different study periods:
G1 - first period grade (numeric: from 0 to 20)
G2 - second period grade (numeric: from 0 to 20)
G3 - final grade (numeric: from 0 to 20, output target)

5 Peer review guidelines

The following guidelines for peer-review should be followed.

5.1 Questions regarding the content

1. Have all the tasks in the assignment been completed? □ □
2. Does the report contain relevant figures and tables? □ □
3. Has all notation been properly introduced and/or explained? □ □
4. Has the model been properly introduced? □ □
5. Are the results properly presented and discussed? □ □

5.2 Questions regarding the report presentation

1. Does the report have:
   • Title, authors, and date? □ □
   • Introduction? □ □
   • Results and/or conclusions? □ □
2. Has the report been proofread? Have language and spelling mistakes been corrected? □ □
3. Are figures and tables:
   • Numbered? □ □
   • Equipped with suitable captions? □ □
   • Refereed to in the text? □ □
4. Is the text divided into paragraphs and well structured with clear and suitable section headings? □ □
5. Is the report easy to read, and understandable without access to the project description? □ □