NEW Instructions for the end of term project. NO LONGER GROUP WORK! NOW INDIVIDUAL WORK.

Note: The instructions remain the same except for some important differences. I summarise them at the beginning and they are in red in the text below.

The end of term project (EOT) is now worth 95% of your final grade. (5% was allocated to the reading week project). **There will be NO EXAM**. The deadline for submission is Wednesday 13th of May 2020 4 pm. **The project will no longer be a group project. Each student will write the whole project alone**. This does not mean you cannot collaborate on coding and talk about it with others (in fact I encourage this and please feel free to use the MS teams I created for this purpose) **but you must write up your own technical and lay reports**. I will answer all questions on the project forum.

Project forum

I will answer:

1. Questions about how to code something (this may be especially relevant to students who have been working alone because they went home early)
2. Generic questions (e.g. If the VIF > 10, should I remove the variable?)
3. Questions about the instructions
4. Questions to clarify definitions of the outcomes/predictors

I will **not** answer:

1. Specific questions: (e.g. should I use the predictor xxxx in the regression with yyyy as the outcome)
2. Is my regression correct?

Submitting work

Each student will upload a pdf of their project. I want a cover page with the words "ST211 Project" followed by the candidate number. Each student will also upload a copy of their R code and the report to Moodle by the same time. Late submissions will be considered on a case by case basis.

The name of your uploaded file should be "candidatenumber_ST211". Your report should be no more than 6000 words long. This does not include the R code or the appendix (details later). Please feel free to make it shorter. As with the reading week project, there will be a technical part and a lay part. The technical part will consist of two separate regression models. One will be a multiple linear regression with a continuous outcome and the second part will be a logistic regression with a binary outcome. The lay part will cover BOTH the linear and logistic regressions.

I will pass the reports through Turn-it-in to ensure that there is no plagiarism.

Working circumstances

Finally, I would like students to submit a short paragraph describing how/where they worked. For example, if you worked alone for the whole project or whether you worked in a group for a few weeks and then alone etc. For example if you had to work in a crowded house and did not get a lot of space to do the work (and were expected to look after children or contribute a lot of time to housework), if you frequently had a poor internet connection etc. I will take this information into

account as I do not want to penalise students because of their personal circumstances at this challenging time. This can be appended to the end of the report. Please be honest.

Questions and Feedback

I will be releasing individual feedback in the form of a set of bullet point comments within 6 weeks of the deadline.

Questions

Please post questions on the project forum (I will ensure it is linked with my email address so I receive notifications of any posts) and I will reply as soon as I can.

Data

For this project, I am giving you data the Next Steps Study (NS) . You are already familiar with the NS as your Reading week project used these data. As a recap the Next Steps (NS), previously known as the Longitudinal Study of Young People in England (LSYPE), follows the lives of around 16,000 people in England born in 1989-90. The study began in 2004 when the cohort members were aged 14, with an original sample of 15,770 people. Cohort members were surveyed annually until 2010, and the next sweep after this was when they were aged 25, in 2015-16. The data I am giving you are not linked to the National Student database.

These are the data as they were originally collected and they still have the complicated names. For example all variables start with WX where X is a number from 1-8 representing the wave in which the information was collected. I have compiled data from waves 1 to 8 excluding waves 3 and 7 on a large number of different predictors. The data dictionary and translator are on the Moodle page. The translator has the year the variable was measured, the school year of the young person (where relevant), the age of the young person, the name of the variable, its type and what it is. The exact details of its coding are in the data dictionary. If you want to know the original data dictionaries, please let me know and I can send you a copy. You will have to search the data dictionary for your variables. Spend some time familiarising yourself with the data. I also recommend that, as you will be running multiple regressions, you think in advance about all the predictors you should take into account for each outcome. Please post in the forum if you have questions or are struggling.

The outcomes

You will be running at least THREE separate regressions. The MLR part will be worth 80% of the final report grade and the logistic regression will be worth 20%. One will consider two continuous outcomes and at least one binary outcome.

Multiple linear regression

For the continuous outcome, you can choose two from:
W8DINCW
W8GROW
W8NETW
W8PUSA
W8NETA

W8QDEB2
which are income/debt related outcomes in Wave 8. It may well be interesting to explore more than one of these (at most 4) and **extra marks may be given for presenting a bigger picture**. Consider carefully whether to include income variables in Wave 8 as predictors as well as outcomes in these regressions.

Logistic regression

The binary outcome can be one of W8QMAFI, W8DWRK or W8QDEB2 . W8QMAFI will have to be turned into a binary variable. I suggest grouping the "Living comfortably" and "Doing alright" categories together into a "Well" category and grouping the other levels together in a "Poorly" category. W8QDEB2 also needs to be dichotomised where any debt could be 1 and no debt could be 0 (other ways can also be considered but must be justified)

Missing/Refused/Drop-out

There are a lot of missing, refusal, don't know etc. values in the data (typically these are coded as negative numbers, -1, -9, -99, -8 etc.) There may also be some NAs where there has been drop-out between subsequent waves. I have removed individuals who are not in Wave 8. Due to drop-out the initial sample has gone from 15,770 to approximately 5,800. At some point in the report *briefly* discuss the possible implications of this reduction on your results. Is the dropout "informative"? i.e. are people with lower income more likely to drop out?

The appendix

As with the Reading Week project, you will probably merge levels of some of your categorical predictors. This information should go in the appendix. Any other data manipulations (e.g. considering complete cases only etc.) should be listed and briefly justified in the appendix. Do not include in the appendix any information about the analysis itself as I will not count it.

Report

You received exhaustive advice on how to write a report for the reading week project. Refer to that document. The word count is longer so you may write more for this report in each section than you wrote for the reading week report. I won't be checking for individual section word counts. In this report you will include discussions of outliers, transformations, interactions, multicollinearity and some validation in addition to what was covered in the reading week project. Refer to the Guide to Model Building at the back of the course notes. If you cover everything that is listed there and in approximately the order suggested there you will definitely pass. I expect you to attempt all the techniques we discussed in the class and to check for all possible problems (collinearity, outliers etc.). If these are not a concern, then it is sufficient to mention that you attempted them and that the results were not interesting.

Group work and the lay report

You should no longer submit the same technical report if you work as the rest of your group! I expect each person to write the TECHNICAL and LAY part of the report ALONE. I will be checking for plagiarism in this part of the report. The lay report can either be divided into a lay report for the multiple linear regression and another for the logistic regression or, if appropriate, the two can be written up as one.

Research question

Because you have a choice in the variables that you select as outcomes and predictors, you need to formulate a research question. This should be clearly stated in your introduction and lay interpretation. I will look favourably on a well formulated research question.