ST300 Individual Summative Project

Aim: The aim is to analyse the data provided with R (other packages not allowed) to explain how the Price-earnings ratio (PE) - a popular metric used to evaluate whether the price of a stock is high/low compared to similar stocks - is related to other variable(s). For example: what can you say about the relationship between PE and other variables? Which variable(s) are potentially more important in determining the PE. These factors could include industry, region of the world, growth rate.

You have a choice of data sets.

- The simplest is to use the data set provided. The data-set *Market_data.csv* consists of cross-sectional data on financial variables across industry sectors, compiled in January 2019. A description of the variables, and a markdown file for cleaning the data is in Moodle.

- Alternatively, you may gather your own data, which is like the one provided but with a more comprehensive set of predictors. The variables must include a mix of continuous and categorical predictor variables with more than 2 levels. Bonus marks awarded for a good data set. Sources of data:

    - The platform **S&P Capital IQ** available online through LSE library is popularly used in the finance industry for gathering company data.
    - **WRDS**. Note on registering. In the Department field, enter ST300.

      If you have registered for WRDS in ST326 you do not need to re-register.

Deadline

The due date is: **12:00pm (noon) Wednesday 20 January 2021**

Assessment

This question counts for 15% of the total mark for ST300. The project is marked out of 100 excluding bonus marks and credit will be given as follows:

Introduction	10
data description and EDA	
Model	40
Variable selection; diagnostic checks and dealing with any issues; are parameter estimates plausible?	
Interpreting the model parameters	10
Understanding the limitations of the model	10
Quality of write-up /narrative	30
Penalty/Bonus	
Excessive figures/tables without discussion	-10
Gathered own data with a reasonable mix of predictor variables	5

The work will be subject to checks for plagiarism.

How to submit your project

Your work should be submitted anonymously under your candidate number. You can look up your candidate number in LSE for you. By the deadline, you must have:

- If you have used the given data: upload your project **with appendix** using the specified Moodle link on the ST300 Moodle. The file should be in PDF format (not zipped). The Moodle upload link will stop working after that time and you will only be able to email me your project
- If you have collected your own data: do the above and upload your data set.

When the submission deadline has passed, a list of candidate numbers for the work received will be put up on Moodle. Each student must check from that list that their work has been received.

Extensions to deadlines for coursework will only be given in fully documented serious extenuating circumstances.

Penalties

Penalties will apply for late submission of either the printed or the electronic version of your project (but will not be doubly-penalised if both are not submitted on time).

Penalties will apply for late submission, at 0.5 mark each day (10 marks total).

There will be 5n marks deducted from the total possible 100 marks for this coursework if either the hard copy or the electronic version is submitted between 24(n − 1) and 24n hours from the stipulated deadline (including working days only).

Getting help

This is an individual assignment. It is expected you apply the lessons of the labs, and you may need to explore outside the classroom for things like coding (Stackexchange.com is helpful), and understanding the meaning of the financial variables.

There are no bonus marks for using tools not taught on the course - indeed you will be penalised if you use a method where it has been incorrectly applied. Do not apply machine learning methods not taught on the course. So be inquisitive, but be careful of straying away from the course content.

You do not need to use everything taught in the course. It may be that you do a regression using material from Weeks 1-5, or you may judge that the model should be a GLM and use materials from Weeks 7-11.