Seminar 6 – week 43

Exercise 1 – Debt (From Exam 2019)

In the seminar, we focus on sub-questions d) to f)

The IMF has collected a database called the Global Debt Database with levels of private and public debt for all countries in the world over the period 1950-2017 (but with quite some missing data). The data are available in Stata format in the file `GlobalDebtDatabase.dta`.

The variables you are going to use are

`ifscode`	IMF code of the country
`country`	Name of the country
`year`	Observation year
`pvd_ls`	Private debt, loans and debt securities, % of GDP
`gg`	General government debt, % of GDP

a) Read the data into R. How many observations are there? How many countries are present in the data set?
b) Select the subset of data from Norway for the period 2000-2017. Define total debt as the sum of private debt (`pvd_ls`) and government debt (gg). Plot the evolution of total debt over the period. You should not spend time on labels etc. Show how the share of private debt in total debt has been evolving over the period.
c) Compute the average share of government debt as share of GDP (gg) for the Nordic countries (Denmark, Finland, Iceland, Norway, and Sweden) for each year in the period 2000-2017). (Ignore the fact that population sizes may differ, and simply compute unweighted averages.)
d) The tab separated text file `countrynames.csv` contains a list of the `ifscode` of the countries in the data set as well as the name of the country in the local language (`local_name`). Import the data into a tibble `countrynames`. Keep the values from 2017 from the GDD data, merge in the local countrynames, and make a scatter plot of the level of government debt (gg) against the level of private debt (`pvd_all`), labelling each observation with the country name in the local language. Make sure the text is possible to read.
e) In the file `countrynames.csv`, country names are encoded in UTF-8 Unicode. Explain what the encoding means, and explain the advantage of using UTF-8 instead of ASCII encoding.
f) Choose the 10 countries with the highest rate of government debt (gg) in 2017, and extract the government debt? (gg) for the period 1990-2017 for these countries.

Exercise 2 – Predicting school performance

In this exercise, we are going to use a sample of Portuguese secondary school pupils. We have data on their final grades and a number of background characteristics. The data can be found as `student-mat.csv` on Canvas – see https://archive.ics.uci.edu/ml/datasets/Student+Performance for a full description of the data.

1) Read the data into a tibble and keep the variables `sex, age, Mjob, Fjob, traveltime, studytime, failures, absences, G3`. Make a variable `grades` with the grades. Convert the text variables to factor variables (`mutate_if` from dplyr may be useful for this).
2) Use `initial_split` from Tidymodels to split the data so 80% are used for training and the remaining 20% for testing. Make sure you construct two separate data frames for the two purposes. Why do we partition the data? Is this the place in the process to perform the split?
3) To use the factor variables in a machine learning framework, we need to convert them to sets of dummy variables. In TidyModles the most efficient way to convert factor variables to a set of dummies is to construct a recipe and use `step_dummy`. Make such a recipe.
4) Train the machine learning models listed below using `fit` from Tidymodels.
   You should keep each of them as a named object (e.g. `ols`) to refer to the models later.
   Train the following models:
   a. An ordinary OLS model
   b. A LASSO model
   c. A random forest
   d. A Gradient Boosting tree (xgboost)
5) Use `predict` to predict grades in the test data you constructed in exercise 4). Construct the prediction error (prediction-true value) for each observation in the test data for the three models. Compute the mean squared error and plot the density of the errors. Which model seems to perform the best?