Predicting student grades using Bayesian linear regression.

Matt O'Reilly, Kordian Pawelec


THE IRISH NEWS

Errors found in Leaving Cert calculated grades system

Around 6,000 students are thought to be affected

Objective: Our objective will be to create a model that can predict grades based on student information. We will aim to use variables like Mock Examination Results, Gender, Study time, Absences, Failures, and Parents Education to predict Final Grade.

Background: 2020 marked the first-ever occurrence of predicted grades for students in Ireland. The Irish government had to provide an estimated mark across all subjects for each student in the country. They got it very wrong! Using a dataset of student grades, we want to build a model that can predict a final student's score from personal and academic characteristics. We found a dataset online with 649 observations and 33 variables, Each row is a student with each column containing a different characteristic.

The primary variable of interest is the overall grade. It can be accurately predicted from the first and second-semester grades since there is a strong correlation between them. However, An accurate prediction of the final grade without semester grades would be much more useful. Since our task is to provide a superior grade prediction model for the leaving certificate we may use the first-semester grade (which in our case is the Mock exam results.)

Proposed Method: We are going to use n=649 secondary school students as our population sample and the experimental units will be each individual student. The response variable will be Overall grade (Continuous variable) while some of our explanatory variables will be Sex (Binary, in our dataset), Age (Discrete variable), Failures (Discrete variable), Parents Education (Nominal variable), Parents Jobs (Nominal variable), study time (Ordinal variable), internet (Binary variable), health (Ordinal variable), and Mocks examination grade. (Continuous Variable)

Scale of Variables: The Overall and Mock examination grades will be measured on a continuous scale from 1 to 100. Sex and Internet are binary variables and so will be measured on a nominal scale. Age and Failures will be measured on a discrete scale. Parents Jobs will be measured on a nominal scale, while Parents Education will be measured on a discrete scale (0 - none, 1 - primary education, 2 - Junior certificate, 3 - Leaving certificate or 4 - higher education) Health will be measure on an ordinal scale (from 1 - very bad to 5 - very good)

Prior choice: we are going to use a normal distribution for β and a half-Cauchy distribution for σ. The posterior is equal to the likelihood of the data times the prior for the model parameters divided by a normalization constant. If we have some domain knowledge, we can use it to assign priors for the model parameters or, we can use non-informative priors: these are distributions with large standard deviations that do not assume anything about the variable.