Using linear regression to predict students’ academic performance

Introduction

An ability to recognize students’ academic performance is now necessary for educators and students. It provides insights into learning patterns and guides decision-making in education institutions on how to improve it. Thus, this research is carried out to identify the factors that affect student achievement. In this project, a linear regression model was used to assess students’ academic performance.

Data Description

In this research, we use Student Performance Data Set retrieved from UCI Machine Learning Repository, which was obtained by Paulo Cortez at the University of Minho[[1]](#footnote-1).

The data is considered to be collected during the 2005-2006 school year from two secondary schools, from the Alentejo region of Portugal. The data attributes include student grades, demographic, social and school-related features, and it was collected by using school reports and questionnaires.

The questionnaire was reviewed by school professionals and tested on a small set of 15 students to get feedback. The final version contained 37 questions in a single A4 sheet, and it was answered in a class by 788 students. 111 answers were discarded due to a lack of identification details. The data was integrated into two datasets related to Mathematics (395 records) and the Portuguese language (649 records) classes. In this project, we will only consider the dataset of Mathematics.

Response variable:

G3: final grade (numeric: from 0 to 20)

Explanatory variables:

G1: first-period grade (numeric: from 0 to 20)

G2: second-period grade (numeric: from 0 to 20)

Address: student’s home address type (binary: ‘U’ - urban or ‘R’ - rural)

Absences: number of school absences (numeric: from 0 to 93)

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

Mjob: mother’s job (nominal: ‘teacher’, ‘health care related’, ‘civil services (e.g. administrative or police)’, ‘at home’ or ‘other’)

The remaining variables were not used because most of them are binary variables with ‘yes’ or ‘no’ and they are not our interest of study in terms of factors.

Who will do what?

After discussing in group, we have decided on the following individual responsibilities:

Yubo Wang has overall accountability for the linear regression analysis in R. This includes assuring proper package to be used, loading data to RStudio and establishing appropriate linear regression models, saving the outputs. Also, write the methodology section in the final report.

Mariana Chen is responsible for analyzing the R outputs, including the parameter estimates in the model and any plots produced, as well as writing the analysis section.

Jason Shao is responsible for writing the introduction section and editing the final report format. Also, he will assist the team members in tracking status and ensuring that the deadline is met.

Zongyang Gao is responsible for writing the conclusion section and providing any further suggestions for future study.

References

Student Performance Data Set page URL: <http://archive.ics.uci.edu/ml/datasets/Student+Performance>. [Accessed 02/11/2020]

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April 2008, EUROSIS, ISBN 978-9077381-39-7.

1. http://archive.ics.uci.edu/ml/datasets/Student+Performance [↑](#footnote-ref-1)