Intermediate Statistics

Matthieu Cisel - CY Tech

September 2021

1 Introduction

In this learning unit, students focus on the practical applications of a set of statistical models with R, namely: the linear model, the logistic regression and survival analysis (Cox model). Regarding the linear model and the logistic regression, they will learn to describe the results of an ANOVA table and of an Odd-Ratios table, respectively. Regarding the survival analysis, we aim at teaching the concept of hazard ratio. Students learn these concepts and techniques through a hands-on approach, and more specifically, they are provided with a dataset on learning analytics drawn from a series of MOOCs. The focus of the analysis lays on learners' engagement in an online course. They will have the opportunity to compare their results with the ones that were obtained in this research work carried out in the mid-2010s. As opposed to learning units on data mining, students benefit from a strong guidance and scaffolfing at each step of the process. One of the focuses of the class is to learn how to design sound figures and captions, and describe the results that are reported in said figures and tables.

This is a R project. You will gain experience in manipulating datasets, perform and present a logistic regression, a survival analysis, and a mixed model based analysis. We provide the raw datasets, just follow the instructions. A template is also given with this document, in order to specify expectations (how to describe an ANOVA table, an odd-ratios for logistic regression, and hazard ratios for survival analysis).

2 Presentation of the context

One of the most striking consequences of Massive Open Online Courses (MOOCs) openness is undoubtedly the high heterogeneity of their registrants, whether we think in terms of socioeconomic status, sociocultural background, motivations, or behaviors. Their engagement patterns are as heterogeneous as their profiles, and the monolithic distinction between completers and dropouts is not necessarily appropriate to describe the diversity of situations.

Most of the time, a large proportion of registrants could still represent a significant part of the course activity despite the fact that they do not complete the course. While these questions have attracted considerable attention from researchers and practitioners lately, few studies have focused on the long-term evolutions of these learning engagement patterns in a given course. Increasing attention is laid on the relationships between these engagement patterns, intentions, or sociodemographic variables. These questions are relevant to both course designers who would like to understand ongoing dynamics and wish to adapt course design accordingly, and to researchers who want to capture ongoing trends at a more global scale. In both cases, even comprehensive studies based on large numbers of MOOC learners.

In this exercise, we propose you to analyze a MOOC that has been organized on Canvas before being on Coursera, the MOOC Effectuation. We addressed the question of the evolution of learners' profiles and course dynamics over time. To what extent have engagement patterns and registrants profiles evolved across iterations, and most importantly, how has the relationship between learners' behavior and profiles evolved over time?

The case study you will analyze here is a five weeks long entrepreneurship course called Effectuation (Professor Philippe Silberzahn, EMLYON Business School), which will thereafter be referred to as MOOC1. It was hosted by a MOOC agency which used the open source LMS Canvas from Instructure.

It was necessary to submit a peer evaluated mid-term assignment and to pass an exam to earn the certificate. In both courses, new course material including quizzes and half a dozen of short videos was made available every week. Variations among iterations were minor. Course designers estimated that completing the course required fifteen to twenty-five hours. Student activity reports, gradebooks and survey responses were downloaded from the platform. Regarding video, consumption, we used a proxy as we considered that the video had been viewed when the page where it was, embedded was opened, regardless of the number of times this page was loaded. We manually removed from, subsequent analyses the videos that were not part of the course strictly speaking, such as weekly introductions or, tutorials. The global activity of the course was defined from the video perspective as the total number of views, without taking into account multiple views, and from the quiz perspective as the total number of submissions, without taking into account multiple submissions.

Participants were asked to fill in a survey at the beginning of the course; response rates ranged from 40 percent, to 60 percent of enrollees. IP addresses were not collected; all available data on countries of residence come from these, surveys; the Human Development Index of these countries were retrieved from U.N data. In both courses, the students who could gain credits by completing the course were excluded from our analyses since they, were not strictly speak-

ing following a self-directed learning approach. They represented a significant contingent, in the case of MOOC2. Participants were categorized based on their level of engagement: those who obtained a, certificate were called "completers", those who submitted at least one quiz or assignment but did not complete, the course were referred to as "disengaging learners"; those who did not submit any quiz or assignment were, referred to as "auditing learners" if they had viewed at least 10 percent of available course videos, and bystanders, if they fell below this threshold.

3 Data wrangling, feature engineering

Patch together the information from usages.effec (all three iterations) and the surveys. If you use R, use merge (base), or preferably full-join functions (dplyr library), and well as rbind, or rbind.fill if need be. Do not forget to keep the information on the iteration of the course when patching together the datasets (using the mutate variable from dplyr). You must deal with the fact that all datasets do not have the same number of columns.

4 Describing behavior in the courses

Participants were categorized based on their level of engagement: those who obtained a certificate were called "completers", those who submitted at least one quiz or assignment but did not complete the course were referred to as "disengaging learners"; those who did not submit any quiz or assignment were referred to as "auditing learners" if they had viewed more than 10 percent of available videos, and bystanders (Anderson et al. 2014) if their fell below this threshold.

Present the proportion of Disengaging learners, auditing learners, bystanders, and completers. This typology of learners is inspired by a paper from Kizilcec et al. (2013) : Deconstructing Disengagement Analyzing Learner Subpopulations in Massive Open Online Courses.

Represent the proportions of learners as a pie chart. The db is called usages.effec1 for the first iteration. When describing the results, use numbers and sentences.

5 Linear Model

5.1 From Student's t-test to two-ways ANOVAs

Group together, for the HDI variable, the High and Medium level to create a new intermediate level. Compare the number of views of videos between genders. Which test should you use to assess whether the difference is statistically significant ? Present the results through a table.

Compare the number of views of videos depending on the HDI of the country of origin. Same questions. Which test should you use to assess whether the difference is statistically significant ? Present the results through a table. What is the difference between the two tests you just used ?

Use Gender, HDI and socioeconomic status as explaining variables (lm command in R, lm(y x1+x2)). Introduce an ANOVA table (anova(model) in R) in your report.

5.2 Model refinement, pairwise comparisons

Update the model, and add an interaction parameter in the it (For instance Gender*HDI in R). Use the summary of the model to see the interaction parameter. Use a stepwise algorithm (step command in R) to assess the performance of various versions of the model (use both forward and backward options). Assess the colinearity of all three independant variables of the last model (excluding interaction parameters). To do that, use a chi-test between HDI and Gender, produce a mosaic plot and propose its interpretation (look for residuals below -2 or above 2). Then, use Tukey HSD, and propose a table, to see the pairwise differences between learners of different socioeconomic status.

Introduce an ANOVA table (anova(model) in R) in your report using at least one interaction parameter. This time, describe it, with p-values and F statistics included in your sentences. Looking for inspiration ?

Check this page : https://mypages.valdosta.edu/mwhatley/3600/oneway.htm

In order to get a better understanding of the issue of pairwise comparisons, we designed a dataset with many continuous variables. Use parwise comparisons with the lm model to detect statistically significant relationships between variables. What variables appear to be correlated ? Include a graph in your report and comment it.

Now apply the Bonferroni correction. What happenned ?

6 Logistic Regression

6.1 Producing an Odd-Ratios table

Use a logistic regression model (glm in R, binary family) to test whether completion, in the course, is linked to the user characteristics that you studied earlier. Make an odd-ratio table. Signal the odd-ratios that are significant in terms of p-value (with stars). Interpret the results (600 words minimum) by providing at least two alternative explanations for how socioeconomic status, or human development index, is linked to completion.

6.2 Poisson models for count data

We used a linear model to study the number of videos viewed, but was it really legitimate ? Plot the distribution of the variable (videos viewed). Produce a qqplot and discuss the homoscedasticity of the residuals. Include the graphs in your report and comment them. Propose a glm model to explore the same questions that you explored earlier, but this time, use the Poisson family.

7 Survival Analysis

You are going to perform a survival analysis on video consumption, in the spirit of what was done by Reich (MIT). https://er.educause.edu/articles/2014/12/mooc-completion-and-retention-in-the-context-of-student-intent. Look at Figure 1. You must reproduce a similar figure based on the data you have at hand. You must reason in terms of proportion of the available videos that the learner viewed. Prepare the data so that they are fit for a survival analysis. Compare video consumption behavior between auditing and disengaging learners, but this time with a survival analysis (and not the linear model like you did earlier). Compute the hazard ratios, and plot the survival curve. Where do you see the most significative drop in terms of video consumption ?