Juan Carlos Ventosa

IST 707 Data Analytics

Homework Assignment 2

Submit Date: 7/14/2020

**Introduction:**

Schools play an important role in preparing children and adults with the knowledge and skills needed to succeed in many aspects of life including career advancement, societal contributions and life goal achievements. How well schools educate therefore become the paramount measure on how students and parents choose schools for themselves or their children. Schools in turn place high importance to educate students to succeed which builds a reputation to not only attract the best students but to also provide suitable candidates to job recruiters at the high school and university levels. A high standard for education and a high graduation rate become the currency for a school’s success and part of the determination for its ranking.

There is an expected association between a school’s standing and the success of its students after graduation. Students who graduate from the highest ranked universities expect to achieve high incomes and career advancement. Employers expect candidates from reputable universities to have the knowledge and skills to contribute to the business’ success. Consequently schools and universities must deliver an education that meets or exceeds the standards the student and employers expect from a school’s repute. Therefore schools and universities must continually improve its methods to educate in order to maintain its standards and meet the expectations of its patrons.

To remain competitive or to raise its standings, schools can employ studies to find ways to improve the proficiency of their students and their graduation rates. Determining key factors that aid or harm the learning process may provide insights to learning or teaching enhancements. Evaluating selection of courses, curriculum, teachers, student population, enrollment practices, technology options and class size offer some suggested areas for exploratory improvements. Analyzing the available sources of information in these areas may discover new methods or practices towards a better education.

**Analysis and Models:**

**About the Data**

The dataset is a summary of student completion rates to the same math course taught at 5 different schools for the current semester. The course has 35 lessons to complete and the semester is about ¾ of the way through. There are a total of 30 records and 8 attributes to the dataset containing a total of 1,601 students distributed into 30 sections.

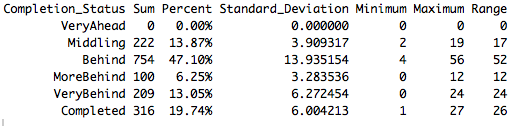
The 8 attributes in the dataset include:

* 2 nominal attributes:
  + Schools - with 5 factors: A, B, C, D and E
  + Sections – ranging from 1 to 13 sections per school distributed as follows:
    - A – 13 sections
    - B – 12 sections
    - C – 3 sections
    - D – 1 section
    - E – 1 section
* 6 numeric attributes recording the number of students who are:
  + Very Ahead – more than 5 lessons ahead
  + Middling – 5 lessons to 0 lessons ahead
  + Behind – 1 to 5 lessons behind
  + More Behind – 6-10 lessons behind
  + Very Behind – more than 10 lessons behind
  + Completed – finished with the course

The dataset required no cleaning with zero missing values and zero duplicated records. One data attribute transformation was required; converting the Sections data type from integer to factor/nominal. The sum of students to a section was added to the dataset as a ninth attribute named Section Totals

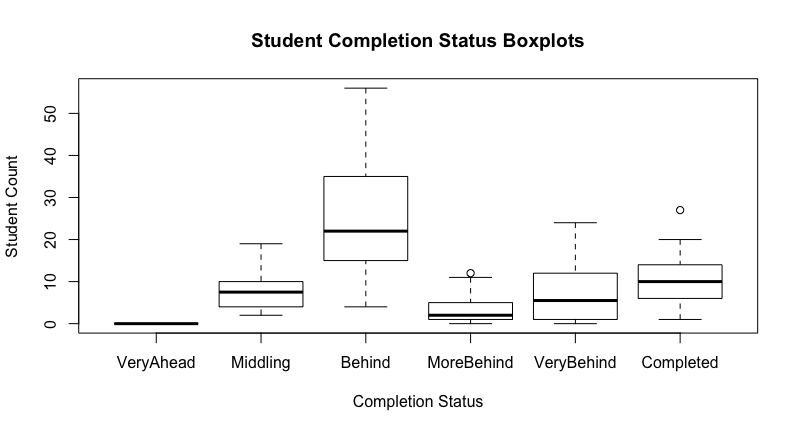
To get a baseline distribution of all the students in the dataset, Table 1 illustrates the aggregated sum, percentage, standard deviation, minimum, maximum and range of student counts by their course completion status. The baseline distribution revealed about 20% of all students completed the course and approximately 14% as Middling or “on track”. The majority (approximately 66%) of students are behind as represented by the sum of Behind, More Behind and Very Behind completion statuses.

**Table 1** Sum, Percent, Standard Deviation, Minimum, Maximum and Range (maximum-minimum) of student counts by course completion status.



Along with accounting for the largest proportion of students (47.1%), the Behind completion status also has the largest standard deviation and range representing an unusually wide spread in student counts by section as compared to other completion statuses. Determining why some sections have very small counts and other sections have very large counts may provide some insights on a section’s overall performance. On the other end of the spectrum, zero students have a completion status of Very Ahead. However, given the semester is about ¾ of the way through and the definition of Very Ahead as “more than 5 lessons ahead”, the zero count may be due to the probability that fewer than 5 lessons are left in the course at the time that the student counts were recorded.

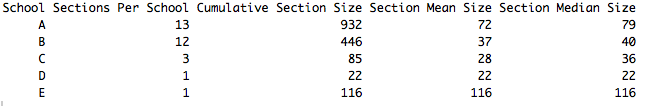
Boxplots as seen in Figure 1 offer an alternative visualization on the distribution of students by completion status. The graph illustrates a similar count, range and median value between students in the Completed, Middling and Very Behind categories while accentuating the nearly triple amount of students in the Behind category. A couple of outliers represented as circles in the boxplots were recorded among students who are counted in the More Behind and Completed categories. The outliers were kept in the dataset during all exploratory data analysis.



**Figure 1** Boxplots representing the median, range, quartile distribution and outliers of students by course completion status.

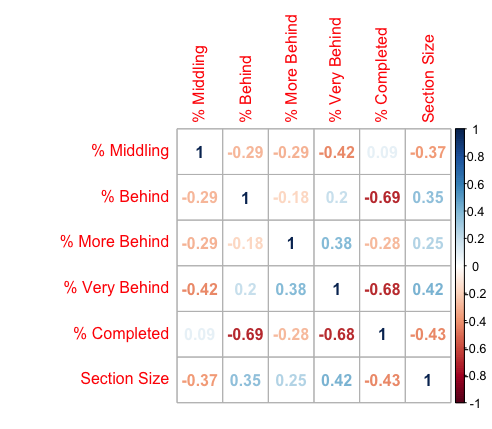
The student population in the math courses also vary by section and school. Table 2 provides a summary of the number of sections per school and the cumulative number of students in all sections by school. There is a clear difference in the student section sizes by school as represented by the section mean size and section median size. With 116 students School E has the largest number of students in a single section. School A also reflects large section sizes with a mean of 72 students per section. Schools B, C and D have smaller number of students per section with mean section sizes of 37, 28 and 22 respectively.

**Table 2** Sections Per School, Cumulative Section Size, Section Mean Size and Section Median Size by School



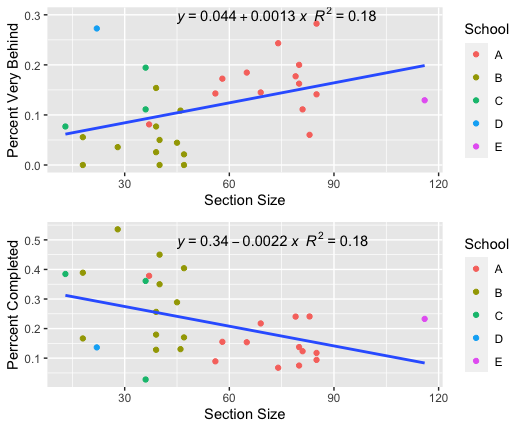
**Models**

Given the differences in section sizes and completion statuses, a correlation may exist between the proportion of students who fall behind and the size of their section. Figure 2 represents a correlation matrix comparing section sizes and students’ completion rates. Completion rates is defined as the number of students by completion status (Middling, Behind, More Behind, Very Behind, Completed) divided by their section size.



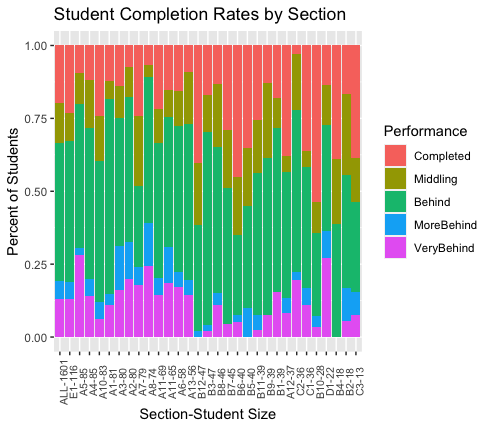
**Figure 2**  Correlation Matrix: Section Size and Proportion of Students by Completion Status (i.e., Completion Status/Section Size). Note: Completion status “Very Ahead” was excluded due to no students in the status category.

The positive correlation between section size and the proportion of students in a section who are in the Behind, More Behind and Very Behind categories suggest larger section sizes will have a larger fraction of students falling behind. The top scatterplot in Figure 3 illustrates this correlation using the proportion of students in the Very Behind category against their section size. Alternatively the negative correlation between section size and the proportion of students in a section who are in the Middling or Completed categories suggest larger section sizes will have a smaller fraction of students who remain on track or move ahead. The bottom scatterplot in Figure 3 illustrates the negative correlation between the proportion of students who completed the course against their section size. Figures 3 uses colors to identify the section’s school and a regression line with equation.



**Figure 3** Scatterplots of percent of students Very Behind and percent of students Completed against Section Size. Color legend identifies the sections’ school and a trendline is added to show the direction of correlation.

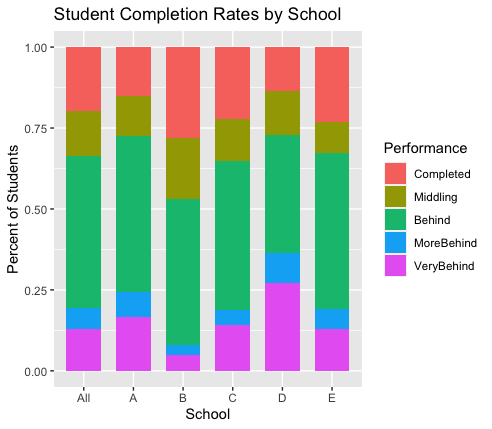
In Figure 4 an area graph is used to illustrate how well each section’s completion rates perform compared to the overall completion rates of all the students in the course. As expected sections with fewer students outperform the overall completion rates among students in the Middling or Completed categories while students in larger sections have higher rates in the Behind, More Behind and Very Behind categories.



**Figure 4** Area Graph of Student Completion Rates by Section. ALL-1601 represents all 30 sections combined with 1,601 students. Remaining X-axis labels represent the school section number with the student size of the section. The X-axis labels are ranked by student size in decreasing order from left to right.

There are exceptions to the rule. Despite having the largest number of students in a section, school E manages to have completion rates on par with the overall performance. Alternatively school C section 2 with the 7th smallest section size had the worst completion rate among students in the Completed category. Interestingly the cumulative effect of these exceptions can change the completion rate expectations of a school’s performance.

Figure 5 uses a stack column chart to compare each school’s completion rates to the overall completion rates of all students. School A with the larger sections does fall in line with having fewer students in the Completed category and more students in the Very Behind category compared to all schools. However, schools D and E provided mix results. School E with the largest section performed on par with the overall performance, and School D with one of the smallest section sizes failed to outperform the overall performance among the Completed and Middling categories. School C only performed on par with the overall performance despite having some of the smallest sections. School B appears to have had the best performance record with nearly half (47%) of its students in the Completed or Middling category and the fewest students falling Very Behind. School B had a mix of small to medium size sections ranging from 18 to 47 students per section, which are below the overall mean section size of 53 students.



**Figure 5** Student Completion Rates by School. All represents all 5 schools combined.

**Results:**

**Correlation**

The correlation between section size and completion rates established a relationship between a student’s likelihood to either fall behind or move ahead in the math course. The key r-values or coefficients in the matrix are the five pairs between the values in the section size column representing the number of students per section and the completion rate rows representing the percent of students within their section by completion category. The remaining pairs in the matrix can be ignored as completion rates are clearly proportional to each other since all completion categories add up to a 100% for their corresponding sections. The five pairs and coefficients of interest are shown in Table 3

**Table 3** Correlation Coefficients between Section Size and Completion Rates

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | % Completed | % Middling | % Behind | % More Behind | % Very Behind |
| Section Size | -0.4289 | -0.3708 | 0.3521 | 0.2528 | 0.4197 |

The correlation between the Section Size and Completion Rate attributes have moderate strengths with no values above or below a ± 0.5. The positive coefficients in the % Behind, % More Behind and % Very Behind categories suggest an increase in section size will increase the percent of students in those categories. The negative coefficients in the % Completed and % Middling categories suggest an increase in section size will decrease the percent of students in those categories. The two strongest correlations or most sensitive to a change in section size are % Completed at -0.43 and % Very Behind at 0.42.

**Single Linear Regression**

Single linear regression models offer a measure of prediction based on the relationship between two variables. The regression line in a scatterplot describes how a response variable changes when an explanatory variable changes. The equation for the line provides the prediction formula for the response variable given the intercept and slope of the regression line.

In the top graph of Figure 3 the response variable, % Very Behind, is shown to have a positive correlation with the explanatory variable, section size. This is represented through the upward slope of the regression line indicating an increase in section size will result in an increase in the percent of students who fall in the Very Behind category. The equation of the regression line, y = 0.044 + 0.0013x can be used to calculate a predicted value for the percent of students who are expected to fall Very Behind (y) given the value of the section size (x). The negative correlation between the response variable, % Completed, and the explanatory variable, section size is represented by a downward slope of the regression line in the bottom graph of Figure 3. The equation of its regression line, y = 0.34 -0.0022x indicates an increase in section size(x) will result in a decrease in the percent of students in the Completed category (y).

The R-squared values of a regression model represents the proportion of the variance for a response variable that’s explained by the explanatory variable as represented by a value between 0 and 1. In both the regression models where % Very Behind and % Completed are the response variables and section size is the explanatory variable, the r-squared values equaled 0.18. Section size therefore only explains approximately 18% of the change in the response variables suggesting other unknown variables may be needed to better explain or predict the changes in the response variables.

**Conclusions:**

The findings of this analysis demonstrated the role a school’s section size has on its students’ completion rates. All 30 sections in the study taught the same math course at five different schools. However, the number of students in each section varied in size and the sections’ completion rates varied as well. Initial observations found that approximately 66% of all students in the course were behind. Those figures varied greatly by section; ranging from 35% to 89% of students falling behind.

By comparing section sizes with completion rates the analysis found that larger sections tended to have a higher a percent of students falling behind while smaller sections contributed to a higher percent of students remaining on track or moving ahead of the lessons. There were exceptions to the rule, and other factors such as the school, professors and the students’ proficiency may also play a role in the completion rates. However, in general larger section sizes resulted in a higher percent of students falling behind.

Based on this finding it is recommended that schools limit the size of their sections to smaller sizes when possible. Achieving this recommendation would likely result in a reduction of students falling behind. However, further exploration of other factors that may contribute to student completion rates should also be considered. Additional analysis may also provide recommended thresholds on maximum or minimum section sizes that would contribute to better student completion rates.