

Assignment: Statistical Tests and Multiple Linear Regression

Part 1: The objective in the first part is to build a profile of successful students and see what (if anything) makes successful students different from the drops. Effectively, we would like to have the ability to determine if any factors contribute to a student being more successful and graduating.

1. This is a real world dataset. I have intentionally left some of the data cleaning for you to perform.
2. Most features in the data file are self-explanatory. Here is the definition for a few variables that may not be clear:
PrevEdCode: Previous Education Code
MinEFC: Minimum Expected Family Contribution
gradFlag: Graduation Flag (1: grad, 0: drop)
3. There are some null values in the file. You don't need to remove those rows. Many functions in R only consider complete observations, i.e., rows with NA's are ignored.
4. When you consider the categorical variables in the Chi Square test, please filter out those levels that have less than 10 observations. Note, that just removing the rows does not eliminate that level. You can use library *forcats* and *fct_drop* function to address it.

What to report:

- a. For the t-test you must create a table reporting the mean of each group along with the p-value. See below as an example.

Feature	Mean (Drop-Group)	Mean (Grad-Group)	p.value
DaysEnrollToStart	57.96	66.81	0.010
AgeAtStart	28.46	26.36	0.000

- b. For the chi-square test you only need to report the p-values in a table.

SexCode	MaritalCode	p
0.491	0	

Note, that the values in the tables above are made up. Don't expect your answers to be the same!

- c. You must write a paragraph summarizing the results as if you were to present it to the stakeholders (Dean, Department Chair, etc).

Part 2: The objective in the second part is to build a multiple regression model that predicts the GPA of students based on the available features prior to their start. As such, you need to only include features in the model that are reasonable. Filter out zero GPA's.

What to report:

- a. You must print the summary of the model, i.e. the p-values, adjusted R², etc.
- b. You must perform complete residual analysis and comment on the LINE assumptions. You don't need to perform any further action in terms of transformation or eliminating influential points, but must explain your observations.
- c. You must write a paragraph summarizing the results as if you were to present it to the stakeholders (Dean, Department Chair, etc).