Student Performance in Exams EDA and Regression

XiaoYuxiang Zhongxiaoda Hezhenlin Wenzihang

Summary of research questions

In this project, we mainly focused on the students' mean score. Therefore, we first come up with two questions.

1. Among those factors (gender, race etc) listed in the dataset, which has the most influence on mean score? Which factors are no that important?
2. Based on the above analysis, could we set up a general regression model to predict the mean score of students depending on their personal information?

Motivation and Background

We are the university students and are interested in getting good grades. These questions are worth analysis and computation since they can give us some suggestions and inspiration.

Based on the result of regression analysis, we can conclude the most influential factors and know the best way to improve the scores. By using our pretrained linear regression model, we can roughly predict the scores of some students given their personal information.

Furthermore, Our predicted scores can also help the teachers to adjust their way of teaching, in order to improve the quality of education.

Dataset

This one-thousand-samples dataset includes scores from three exams and a variety of personal, and economic factors that have interaction effects upon them.

Given is a table illustrating the detailed information of the dataset.

Gender		Race/ethnicity		Parental level of edu.		Lunch		Test preparation course	
Female	52%	Group C	32%	Some college	23%	Standard	65%	None	64%
Male	48%	Group D	26%	Associate's degree	22%	Free/reduced	36%	Completed	36%
		Other	42%	Other	55%				

Table 1: Variables & Proportions

The resource could be downloaded from here.

URL: http://roycekimmons.com/system/generate_data.php?dataset=exams&n=1000

Methodology

To find which factor or variable influences students' grades most, a linear regression model might be a great choice for us. First thing first, variables in this dataset should be checked. Apparently, apart from the three scores, the rest of the variables are all categorical. Usually in a linear regression model, a categorical variable needs to be transformed into a dummy variable. Thankfully, R can automatically achieve this.
The expected formula is :

$$\hat{Y} = \hat{\gamma}_0 + \hat{\alpha}G + \hat{\beta}_1 R_1 + \hat{\beta}_2 R_2 + \hat{\beta}_3 R_3 + \hat{\beta}_4 R_4 + \hat{\delta}_1 P_1 + \hat{\delta}_2 P_2 + \hat{\delta}_3 P_3 + \hat{\delta}_4 P_4 + \hat{\delta}_5 P_5 + \hat{\varepsilon}L + \hat{\zeta}T$$

In the original dataset, there are 2 categories in gender (G), 5 in race/ethnicity (R), 6 in parental level of education (P), 2 in lunch (L) and 2 in test preparation course (T). But in the formula, the number of categories in each dummy variable drops by 1.

With statistical models such as linear regression, one category of each dummy variable needs to be excluded (in R, by default, the first will be excluded) to avoid the predictor variables being perfectly correlated. To better fit customary order, in this dataset, the excluded categories of dummy variables are reconstructed :

Gender : male;
Race/ethnicity : group A;
Parental level of education : high school;
Lunch : free/reduced;
Test preparation course : none.

$\hat{\gamma}_0$ in the above formula represents the estimate of the mean score of students exactly satisfying all the excluded conditions. The rest of categories respectively form each of remaining terms in the formula. For example, G stands for *gender: female*, R_1 stands for *Race/ethnicity: group B*, L stands for *Lunch: standard,* etc.
If a student is from group B and has standard lunch, while sharing the same remaining categories as the excluded ones, then when applying this formula to estimate his mean score, we just let R_1=1 and L=1, and let the rest of R, P and T equal to 0.

It is worth noting that the final formula we get might be different to the above one. That is, some terms may be removed to promote the model. To judge whether a variable should be deleted in a linear regression model, a t-test will be applied on each of them. The hypotheses are like:

H_0 : the estimate of this variable is significantly 0
H_a : the estimate of this variable is significantly not 0

If we apply a significance level of 0.05 which is quite common in daily life, then any P(>|t-value|) < 0.05 will reject the null hypothesis. In other words, if P(>|t-value|) is less than 0.05, then we can be confident about this variable having something to do with the response variable or in short, this variable is credible. If a variable isn't credible, simply removing it usually works but in some cases, might make the model worse.
Therefore, every time we upgrade the model we must check the new one to see whether the previous operation is reasonable or not. This could help us to filter the selected variables. All variables in an ideal linear regression model are expected to be credible, which is what we try to realize. Once we get the *p-value* and other output, among which variable is influential could be determined, thus the first question can be answered.

For the second question, before we build a general regression model, the data need to be reduced to exclude the irrelevant variables. Specifically, those with *p-value* > 0.05 would be reduced to simplify the model. Additionally, regarding the collinearity, the variable 'race' and 'parental level of education' should be analyzed separately, for the underlying relationship between them.
The reduced data is then randomized and divided into training and testing data. In our approach, we use two thirds for training and the rest for testing. To check the validation of our model, a plot of 95% confidence interval is generated and a close look at the adjusted R squared imply whether the mean score of students could be explained by the variables in our model. When the R squared is large enough, indicating a strong relationship between the independent and the dependent score, the model is proved to be convincing.

Collaboration: None

Reflection

The theoretical statistics and programming skill are tightly combined due to this analytical project. Through the exploratory data analysis, data cleaning is of vital importance to lead the following analytical process, which is also a common weakness while learning statistics. Our project comes up with the difficulties while seeking for the most influential factor among various of variables.
The skills to handle with raw data processing and select for accessible variables may greatly enhance our analysis ability. Sometimes adding simple data cleaning task (e.g. provide not-well-processed data) in laboratory exercise is operatable to establish the data cleaning skill.