Title: Descriptive analysis of the effects of Cultural cognition and racial background on the performance of a student

Samrawit Kebreab Tekie

4/26/2021

Introduction

The dataset that is being used for this project depicts a sample taken from a student population of a university. It represents a section of freshman students, and the data was collected for a project that envisioned to assist in identifying the need for an early intervention for students after their first exam. The office of enrollment course registration data was used to collect the information.However, refining manipulation had been done by the analyst of this project before using it for this project. The prior manipulation of the data is not included in this project report.

The data set have different information and fields with major categories including demographic, academic plan, course workload, and academic history based on SAT scores. However, for this project the targeted fields are “race” (as a proxy for background, economic status, and cultural cognition),“Diagnostic Exam”, “Exam One”, “Module 1”, “Academic Plan”, and “total course load”. Here in this report the main focus of the study is to investigate the effects of cultural cognition and racial background on a students performance. We try to evaluate different relationships and try identify if there is a reasonable pattern that can describe the relationships between these social factors and academic achievement. To that end, this project uses exam results from one of the core courses in STEM field. Diagnostic Exam: a test given to students on the first day of class to check and caliber their preparation and prior background for the course material.

Module 1: a test administered after the first unit of the course material. Exam One: a test administered after three units of the course material.

Hypothesis: Null: The background and cultural cognition of a student affect student’s preparedness for college and success in college. Alternative: A student’s preparedness for college and success in college are not related to background and cultural cognition.

Load all the necessary packages

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
import seaborn as sns
import scipy.stats as stats
import matplotlib.pyplot as plt
import warnings
from statsmodels.formula.api import ols
%matplotlib inline
```

Data Processing

The BIO101_project was uploaded to the dataset repository and got loaded into the jupyter note book using the *pd.read_excel()*

Project=pd.read_excel("../BIO539/BIO101_project.xlsx")

Data Cleaning

As mentioned above the author had done a pre-clean up of the data in the excel before importing for this project. During that clean up the data was made unanimous and unnecessary columns were dropped. Once the data was loaded in python further data clean-up was necessary to do the work needed for this project including changing long field titles for columns (eg. "Exam One" to "Exam1", "Diagnostic Exam"" to "DigEx". Also, to make statistical operations go smoothly null values were removed from the data set particularly from columns 'Exam1' and 'DigEx' using *pandas dropna* method.

Data Analysis

Research question 1: this study is trying to address is the existence of a meaningful relationship between a student's background and the student's preparation for college level study. To that end we attempted to investigate if there is a significant pattern connecting ethnicity and diagnostic exam fields in the sample using descriptive statistics where background is represented by ethnicity as a proxy and preparedness test by the results of the diagnostic exam. To investigate the relationship between students' background and their performance in the course the *groupedby* method was employed on the data 'Ethnicity' attribute of the data set. The *.agg* method was utilized on the grouped data to get statistical summaries mean and standard deviation of the column 'DigExam'.

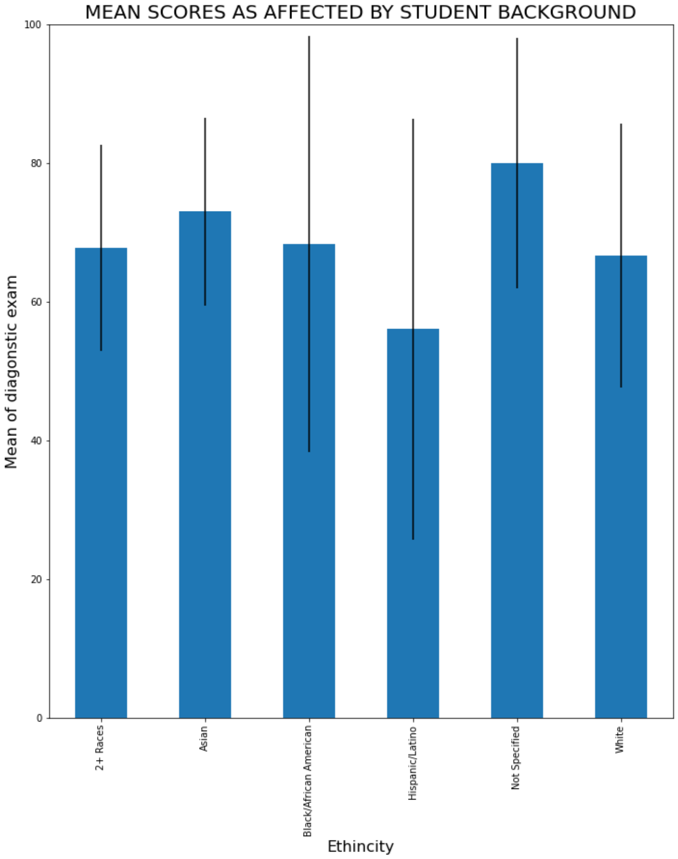

Fig.1. The mean values of DigEx by Ethnicity with standard deviation bars

After aggregating mean and standard deviation, a bar graph drawn out of the evaluated data.This graph shows 'Effects of background on student educational progress. where x-axis represents Ethnicity and Y-axis Mean of diagnostic exam.The Comparable mean values and the overlapping standard deviations indicate that there is no much disparities between the aggregates as a group.

Now, taking the analysis one step further we tried to compare the gap within a group. In order to do that the we tried to see the minimum and maximum scores in the DigEx within Ethnicity.

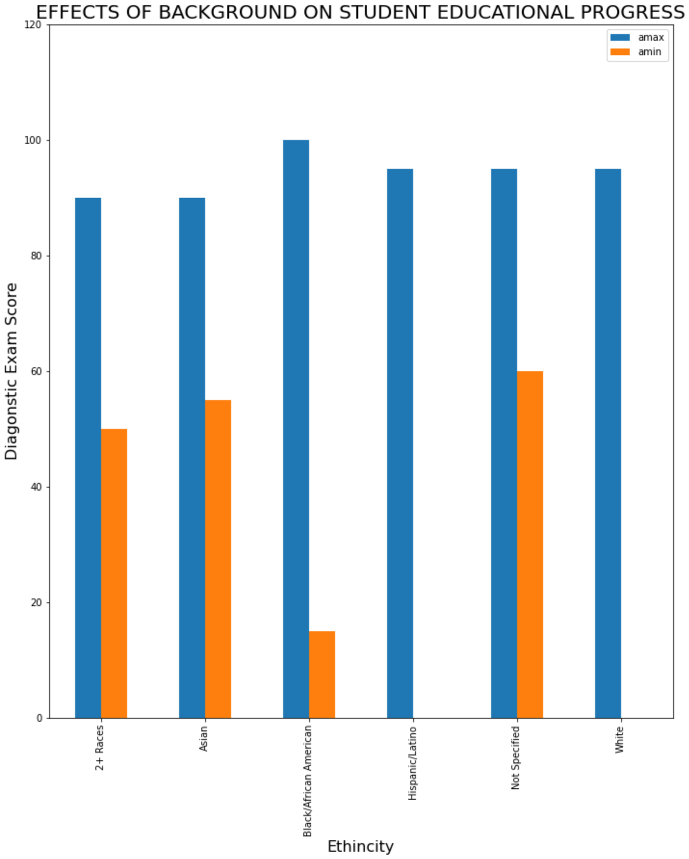

Fig.2. The maximum and minimum values of DigEx by Ethnicity

As we can see in Fig.2, the general trend of the gap within Ethnicity is comparable for Asians, Mixed race, and not-specified. The gap within African Americans shows a larger disparity in preparation for college as the maximum and minimum are at extremes. Based on the results shown below there is a relationship between a students background and preparedness for college education.

Furthermore, the study attempted to investigate the robustness of the above indicated relationship between background and preparedness for college by expanding the background proxies toward first generation and gender fields in order to cover all the subtle parts of the cultural cognition elements that play part on a

student's background beyond ethnicity.
A linear regression approach was used to test this effect of cultural cognition on student's preparation for college level study. Using gender and first generation fields as predictor variables to explain the diagnostic exam results the linear regression model was run as follows. A linear regression analysis was done to describe the results of Diagnostic exam through the effects of gender and first generation interactions. Here a formula approach from the stats models is used, which uses *C()* to indicate a categorical variables.The formula assigns a dummy value to a categorical variable and uses it in the regression. This code was adopted from https://stackoverflow.com/questions/50733014/linear-regression-with-dummy-categorical-variables and modified to fit this project.

```
Intercept 6.833949e-38
C(GENDER)[T.M] 9.077231e-01
C(FIRST_GEN)[T.Y] 7.860944e-01
dtype: float64
```

Table 1. The P-values of the regression analysis for Diagnostic Exam being affected by gender and first generation attributes of the data.

The high p-values in this recessional analysis do not show support to the null hypothesis that there is a correlation between a students preparation for college and student's background and cultural cognition explained by gender and first generation fields in the dataset. In the context of background and cultural cognition we have also seen the effects of ethnicity as factor. Therefore, from our analysis it could be deduced that a students preparation as tested by the diagnostic exam shows that there is a strong relationship to the students background and cultural cognition interms of which ethnicity they belong to regardless of their gender and being first generation or not.

Research question 2: Could this influence of cultural cognition and background influence the progress/success of a student in college? Can this relationship be used as predictor of a student's academic success when used with other initial test measures? In order to investigate this inherent connection between a student's success and background the author used the fields ethnicity, gender, and first generation to describe cultural cognition and background, while the results of Module 1 (Mod1) was used as a predictor of success to indicate as to how results of Exam One are reflected.

```
Intercept 7.616668e-07
C(GENDER)[T.M] 8.217939e-01
C(ETHNICITY)[T.Asian] 4.502284e-01
C(ETHNICITY)[T.Black/African American] 5.108928e-01
C(ETHNICITY)[T.Hispanic/Latino] 4.438783e-01
C(ETHNICITY)[T.Not Specified] 1.048524e-01
C(ETHNICITY)[T.White] 9.829024e-02
C(FIRST_GEN)[T.Y] 6.126527e-01
Mod1 2.280686e-03
dtype: float64
```

Table 2. P-values of interaction analysis between Exam one results and Ethnicity, First generation, and Gender attributes using linear regression as an approach.

The p-values greater than 0.05, which are exhibited for fields ETHNICITY, GENDER,and FIRST_GEN show that they have no significant relationship in affecting the students outcome in Exam1, while the 0.002 p-value for Module 1 test scores represents that the influence of Mod1 results on the students Exam1 results is statistically significant. When connected with our results for research question number one this indicates that a student's progress and success most likely will be influenced by the student's preparation as attributed to background and how that manifests in their initial exam in college.
A Residual test (not shown here, see code book) was performed for the regressions model and the model was found to satisfy all assumption for residual requirements. The statistical summary of the mean show that the model has a mean close to zero across the distribution, the vertical spread of the points is approximately

constant between -50 and 50, appears to have a normal distribution where points are clustered close to zero on the y-axis of the graphs, and there is no relationship between the input variables and the residuals showing that variables are independent of one another. This suggests the robustness of the model.

Research Question 3: Does a student's envisioned career path (described by program of study) have an influence on students preparation for college and success in college? Would this relationship change (defy) the effects of cultural cognition and background.
Using descriptive analysis of primary program with Exam1 results in its aggregated evaluation only shows the naturally expected difference between general groups of natural science and humanities.The details of distinction between students among programs becomes more pronounced once we introduce ethnicity and first generations fields as predictors (see plots in code book). The program choice difference between students of different background shows a distinct pattern. Most of all there is higher number of first generation students scoring higher points in the Exam1 for students identified as minority groups, which is contrary to the results and predictions of the two research questions we evaluated in this project so far. This difference explains that the amount of work put into and the preparation made for college has correlation to the students envisioned career path. It is observed that students of minority groups both first generation college students and non-first generation who are in the Environmental and health sciences scored higher than the other group of programs.The scores tend to be lower than the general trend with students of non-first generation college students who are in the humanities programs.

Research question 4: If cultural cognition and background drive a students success in college and career path choice, therefore it is to be expected that a students result in the Diagnostic exam and Exam1 to remain more or less similar. To test this assumption, a students paired t-test was conducted on the fields DigEx and Exam1.

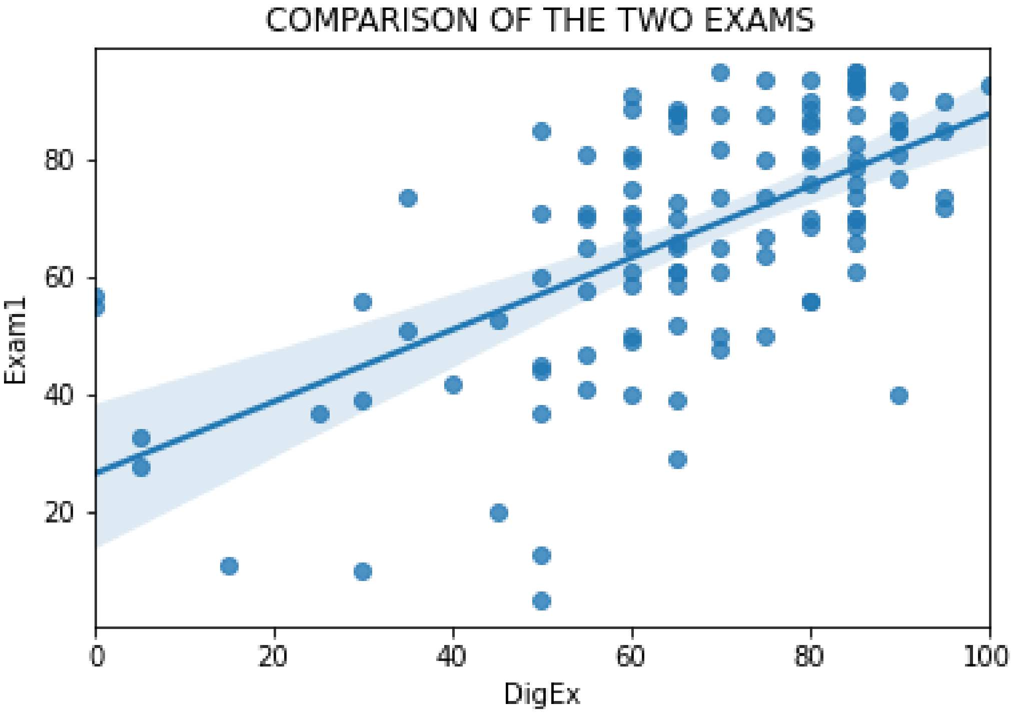

Fig.3. Regression plot for interaction between Diagnostic exam scores and Exam One scores.
Along with the regression plot an independent t-test was performed to test if the two fields Exam1 and DigEx are statistically different from one another. The t-test result shows that two fields are significantly different from each other. Therefore, there are other factors that played into the change of results from DigEx at the beginning of the semester to the Exam1 results almost mid point of the semester. Therefore, we tried to look into if there is a relationship or pattern between the results in the diagnostic exam and the

changes exhibited by students in exam one. A linear regression model (see code book) was used to describe Exam1 via a single predictor DigEx. influencing them.

```
Intercept 1.406149e-06
DigEx 5.692343e-13
dtype: float64
```

Table 3. P-values of regression model expressing Exam One results as predicted by DigEx results.

The linear equation Exam1(DigEx)=26.5742+0.6131DigEx, has a p-value very close to zero showing a strong evidence against the null hypothesis for this question, which is Exam1 and DigEx are results of similar pattern. The results of this analysis conclude that students results shown in diagnostic exam and exam one have different factors

conclusion

The effect of cultural cognition and background on a student's preparation to college is much pronounced as evidenced in question one of this study, while different factors play into a students success in college as evidenced in questions through two to four in this study. Therefore, it needs more rigor statistical analysis and models to predict a students success in college based on the factors influencing the students preparation, level of work in college. This analysis of complex factors is beyond the scope and time of this project.