College Motivation Data Analysis in R (Descriptive)

Richard Murad, Kennesaw State University
Faculty Advisor: MinJae Woo, PhD

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING
School of Data Science and Analytics

ABSTRACT

The goal of this project is to use descriptive analytics to compare variables/groups regarding high school students in Indonesia. 1000 students were sampled and asked questions that include information on their grades in various subjects, their scores on standardized tests, their parents' level of education/occupation, and their demographic characteristics, such as gender and ethnicity. The dataset also includes a binary variable indicating whether the student went on to attend college after high school. With this information, we can use the college interest categorical variable to observe the difference in average high school grades between students interested in college and students not interested in college. A t-test will be used to see if the difference is significant or not. Another great hypothesis test is to see whether the parents' decision to go to college influences the child's decision to go to college. A chi-square test of independence can show us whether these two variables have a relationship or not. Lastly, a multiple linear regression model will be generated to see if the average monthly household salary and the type of institution will influence the students' average high school grades.

INTRODUCTION/EXPECTATIONS

Our variable choice consists of logical approaches that researches might find interesting. In our t-test, we expect the students that are interested in college to have a higher average high school grade than those who are not interested. This makes sense because if a student is interested in college, then they are most likely doing better in high school so they can have more college opportunities. We also expect the parents' college decision to influence the child's college decision. Since parents took that route, it is common for them to push their kids to take the same route. Our regression expectation is the most straightforward; of course, having more income in the family will allow more/better resources for the student, and that typically will help them score higher in their studies. As far as institution type, it makes sense for academic high school students to have a higher average grade since they are typically looking to get into college compared to vocational students, who are looking for a certificate to qualify them for a job.

DATA CLEANING PROCEDURES

For these 1000 students, 6 variables were selected to run analyses on: (1) Parent College Decision, (2) Average High School Grade, (3) Both Parents' Salary (in USD), (4) College Interest Level, (5) Student College Decision, and (6) Type of Institution.

Data Cleaning Steps

- **Variable Creation:** Converting the parents' salary currency from Indonesian to US dollars (USD = IDR / 14,500) and creating the new variable using mapply function
- **Subsetting Variable:** Five different categories of college interest variable subsetted into two categories: "Interested" and "Not Interested"
- **Outlier Removal:** Two outliers removed in "Interested" group to satisfy normality assumption of t-test
- Removed the 'House Area' independent variable in our regression model due to homoscedasticity

Figure 1: Boxplot Distribution

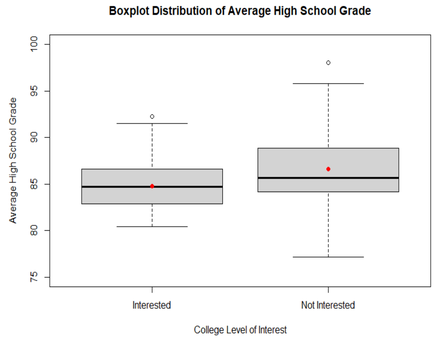

Figure 2: Output of T-Test

Welch Two Sample t-test	
95% Confidence Interval	[-2.797,-0.793]
df	136.05
Mean of Interested Group	84.808
Mean of Non-Interested Group	86.603
t-value	-3.542
p-value	0.0005

Figure 3: Stacked Bar Chart

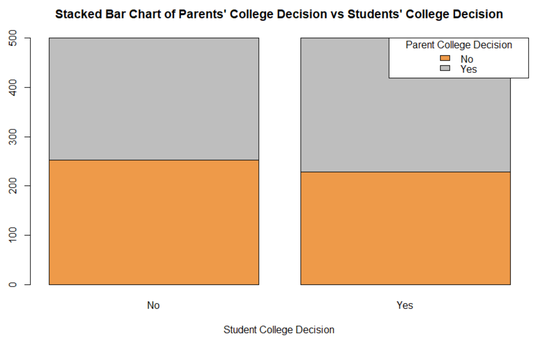

Figure 4: Residual Scatterplot

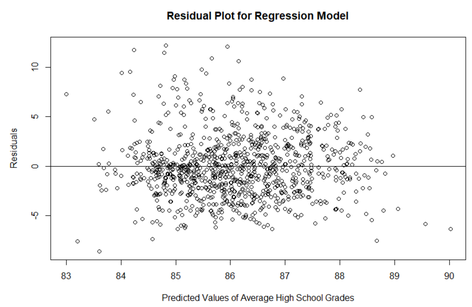

Figure 5: Regression Table

	Coefficients	Standard Error	t-value	p-value
(Intercept)	81.4016	0.4407	184.722	<0.001
Parent_Salary_USD	0.01131	0.0011	10.687	<0.001
type_school	0.8172	0.2089	3.911	<0.001

ASSUMPTION CHECK

Unpaired Two Sample T-test: P-value for var.test() is <0.001; concluding that the variances aren't equal. Satterthwaite will be used here. Both variables fail to reject null in the Shapiro-Wilk test, indicating normality.
Chi-Square Test of Independence: Independent outcomes and expected frequency count is greater than 5.
Multiple Linear Regression: Parents' salary has a 0.31 r-value which indicates a weak positive linear relationship. Type of school has a very weak positive linear relationship (r = 0.075), but it is a binary variable so it will be linear. Residual plot does not show any sign of variance change throughout the line at y=0. Residuals seem to be normally distributed. VIF of both independent variables are below 2. This is a cross-sectional study, so the residuals are assumed to be independent.

CONCLUSIONS/INTERPRETATIONS

Unpaired Two Sample T-test: P-value = 0.0005, Reject Null: students that are not interested in college have a higher average high school grade than students who are interested in college.
Chi-Square Test of Independence: P-value = 0.1454, Fail to Reject Null: there is no association between the 'Parent College Decision' and 'Student College Decision' variables. In other words, there is no significant difference in the proportion of parents who did go to college versus parents who didn't go to college between students who did go to college versus students who didn't go to college.
Multiple Linear Regression: P-value of model = <0.0001, the overall model is significant: 10.78% of the variation in the students' average high school grades are explained by the monthly salary of both the parents and the type of high school the students went to (r^2 = 0.09407). Both predictor variables are significant (p-value <0.001): for every one-hundred (US) dollar increase in the monthly salary of both parents, there is approximately a 1.13-point increase in the student average grade, given that all other variables are held constant; this association is statistically significant. As compared to vocational students, being an academic student is associated with an increase in average high school grades by 0.82 points, given that all other variables are held constant; this association is statistically significant.

DESCRIPTION of USED VARIABLES

parent_was_in_college: Did at least one parent go to college? Yes or No
average_grades: Average high school grades
Parent_Salary_USD: Monthly salary for both parents in USD (converted from Indonesian Currency)
interest: student level of interest to go to college
In_college: Did the student go to college? Yes or No
type_of_school: High school institution type. Vocational or Academic

USED R PACKAGES

Car: Provides various tools for regression modeling.