Teaching Note

PROFESSOR KWAN'S QUANDARY

Amanda Ji, Christopher Kwan, Cindy Nguyen, Ibrahim Rana, Kennedy Confurius, Valay Shah and Yuxuan Zeng wrote this teaching note as an aid to instructors in the classroom use of the case Professor Kwan's Quandary. This teaching note should not be used in any way that would prejudice the future use of the case.

Version: 2020-12-14

SYNOPSIS

Professor Kwan, a local public school math teacher, has always been a strong advocate for equal access to education. As a teacher, he always tried to make sure his students all had equal opportunities. As he read his morning paper, he saw an article on how education expands disparities amongst different ethnic and socioeconomic groups. This made him wonder whether these disparities were present amongst his students and whether some students had inherent advantages over other students when it came to academic testing. Although policies like affirmative action have been put in place to rectify these disparities, Professor Kwan wondered whether this was effective. He gathered his students' reading, writing and math test scores from the previous year along with information on external factors regarding their parents' education level, their lunch status, their ethnicity and the completion of test preparatory courses. He hoped to use this data and determine whether a disparity exists amongst his students and what external factors cause the biggest disparity. If Professor Kwan finds differences in test scores amongst different groups of students, he hopes to determine how he can level the playing field to help his students succeed as they progress through their educational journey. Furthermore, Professor Kwan wonders whether his findings can be applied to education at large and if his findings could have widespread implications.

LEARNING OBJECTIVES

- Use probability laws to gain insights on the effects of various external factors on student test scores
    - Model and visualize probabilities
- Use regression to determine the correlation and relationship between various external factors and student test scores
    - Practice creating dummy variables for categorical variables
    - Conducting regression analysis in various analytical tools (R and Excel)
    - Practice interpreting regression results
- Learn the importance of statistics and apply learnings from statistical models to real-life issues and possible solutions

POSITION IN COURSE

This case can be used in business statistics courses/modules at the undergraduate/graduate level and can be used as a case to further learnings on the topic of linear regression.

RELEVANT READINGS

"Admissions Statistics." *Harvard College*, college.harvard.edu/admissions/admissions-statistics.

Hartocollis, A. (2020, November 16). Harvard Victory Pushes Admissions Case Toward a More Conservative Supreme Court. Retrieved November 28, 2020, from http://www.nytimes.com/2020/11/12/us/harvard-affirmative-action.html?auth=link-dismiss-google1tap

Peter C. Bell and Gregory S. Zaric, "Predictive Modelling," chap. 5 in Analytics for Managers with Excel, 1st ed. (New York, NY: Routledge, 2013), 131-165

ASSIGNMENT QUESTIONS

1. Is there disparity amongst the grades of different groups of students, as a result of external factors?
2. To what extent do external factors impact a student's academic performance?
3. How does the preparation course make an impact on scores of students in need? If student X (whose score is 65%) took the prep. course, what is the probability that he will obtain at least an 80% on the next test?
4. Professor Kwan wants to consider all factors affecting student X's math scores. Student X is a male belonging to ethnic group B, has a modified lunch plan and their parents' highest level of education was a bachelor's degree. Predict Student X's math score both before and after taking the preparatory course.
5. If you were Professor Kwan, would you change anything in your classroom/teaching based on your findings? If so, why and how would this help further affirmative action in education?

TEACHING PLAN

This class will begin with a discussion surrounding initial thoughts on affirmative action and admission practices (15 to 20 minutes). This discussion will serve as a baseline for statistics in college admission processes and help students connect their findings from the case to the real world. Much of the class will be spent reviewing the case questions and implications, ending with a quick summary of learning points to conclude the class (60 to 70 minutes).

Discussion Point	Time (Minutes)
Introduction	5
Assignment Question 1	15
Assignment Question 2	20
Assignment Question 3	15
Assignment Question 4	5
Assignment Question 5	5
Conclusion	5

INTRODUCTION

All referenced exhibits in this teaching note refer to the attached excel file titled “Exhibits”.

Begin the class with a short discussion of the case by reviewing the problem at hand, the decision maker and what factors are being taken into consideration. Ask for initial thoughts on admission practices and whether students think this is an important topic to explore through data analysis.

ANALYSIS

Assignment Question 1

Is there disparity amongst the grades of different groups of students, as a result of external factors?

Start the conversation by asking students what analytical tool they used to determine whether there is disparity amongst different groups of students based on the following external factors: gender, race, parental level of education, completion of test preparatory courses and lunch status.

If students mention regression, note that this will be discussed upon further analysis, but we are looking for a preliminary analysis first.

Summary Statistics

To begin, students can evaluate the summary statistics for each subject. Looking at the scores across the three subjects, it is apparent that students performed slightly better in reading and writing than they did in math (Exhibit 1, Q1 tab).

Conditional Probability

Students can then use conditional probability to observe how student test scores change given different characteristics.

Discussion Point: Ask students the difference between using conditional probability and regression for analysis. The key distinction between conditional probability and regression is that regression will reveal the relationship between dependent and independent variables based on correlation, whereas conditional probability will return the probability of an event given a known characteristic. Therefore, conditional probabilities for different groups of students can provide a preliminary assessment of whether there is a disparity amongst student test scores.

Remind students that conditional probabilities are calculated using the following formula:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Conditional probability analysis can be done in excel by segmenting the data in the student file based on the various external factors and using the =COUNTIF() formula to determine the probability of achieving different average test scores given various characteristics. The conditional probabilities can then be plotted to visualize the probability of achieving different scores amongst different groups of students. The analysis for various conditional probability scenarios is as outlined below:

- *Conditional Probability Results for Gender*
  The conditional probability analysis for test scores given gender indicates that female students are more likely to score higher than their male counterparts. The difference in mean scoring between male and female students is approximately 3.9% and female students are 12% more likely to score above 70%, while male students are 8% more likely to score below the 70% (Exhibit 2, Q1 tab).
- *Conditional Probability Results for Ethnicity*
  The conditional probability analysis for test scores given different ethnicities shows that students in ethnic group A have a much lower probability of scoring well compared to those in ethnic group E. In fact, the probability of a student scoring above 80% in ethnic group E is more than two times that of students in ethnic group A. Aside from the clear difference between these two groups, overall, the ethnic groups C, D, and E have higher chances of achieving an above average grade in comparison to ethnic group A and B (Exhibit 2, Q1 tab).
- *Conditional Probability Results for Socioeconomic Status*
  The conditional probability analysis for test scores given different socioeconomic statuses is done using various factors that may be indicative of socioeconomic status, including parental education level, lunch status and completion of test preparatory courses. Based on the analysis, students whose parents completed higher education can provide standard lunches for them and/or can afford test prep are more likely to achieve higher grades when compared to students who do not have these luxuries.
    - Students with the standard lunch were two and a half times more likely to score above 80% and were eight times more likely to score above 90% compared to students who received free or reduced lunches (Exhibit 2, Q1 tab).
    - A student that took preparatory courses is 6 times more likely to score above 90% than those that did not. However, it should be noted that the mode score for both groups of students is very similar at approximately 65%. This indicates that the courses enable students who score well to score even higher but impacts students who already score low less (Exhibit 2, Q1 tab).

<u>Discussion/Conclusion</u>

Discussion amongst the class can be prompted by asking students what they can conclude based on the relationships we have seen between test scores amongst different groups of students. Based on this analysis students could make the following conclusions:

- There is a clear disparity between average test scores amongst different groups of students based on gender, ethnicity and socioeconomic factors.

Students may also wish to discuss possible reasonings behind the results from this analysis. A non-exhaustive list of the possible explanations for the disparity seen is outlined below:

- Female students may have higher average scores than their male counterparts due to the scientific fact that females mature faster than males, thus, girls perform better than boys on average at this age.

- Females having higher average linguistic scores than males could be partially explained by scientific evidence that females tend to excel in verbal and written abilities, while males tend to outperform females on visuospatial and quantitative ability.[1]
- The gender difference may be caused by gender stereotypes, since mathematics is perceived as a male-dominated subject and writing is perceived as female-dominated discipline.
- Disparity amongst different ethnic and socio-economic groups may be a result of institutionalized prejudice affecting the opportunities and resources available to different groups of students.

Assignment Question 2

To what extent do factors such as gender, ethnicity and socioeconomic status impact a student's academic performance?

Start the conversation by asking students what analytical tool they used to determine the extent to which external factors affect student grades.

Move into the regression analysis by asking students what analysis we can do to determine the relationship between each external factor and students' grades. Note that the regression can be performed in either R and/or Excel. When using R, students merely change the variables into factors and R will automatically categorize non-numerical variables accordingly without the need to manually assign dummy variables, therefore the first step in this process applies only to Excel users.

Conduct regression analysis with students by going through the following steps:

- **Converting categorical variables to dummy variables (Excel only)**

  Start by asking students how they accounted for categorical variables in their regression and what formulas they used to do so. Students could use an IF( ) or COUNTIF( ) formula in excel to change all categorical variables into binary dummy variables, remembering that n-1 dummy variables are required for categorical variables with n unique categories. For instance, there are 4 dummy variables for the 5 races with Race E being the base. If a student is race A, then we will assign 1 in the column for race A and 0 under all the other race columns. If all columns are 0 then the student is of race E. The dummy variables are as shown in Exhibit 3, Q2 tab.

  Discussion Point: How did students decide how to convert each categorical variable into dummy variables and were there any categorical variables that were particularly difficult? Discuss any differences between how students created dummy variables and whether there is a difference between variables created in Excel in comparison to those automatically generated by R.

  Note: Students may try to convert categorical variables into consecutive numerical values as opposed to dummy variables. This method is not advised and incorrect unless the relationship between each categorical variable and the dependent variable is known and can be converted into a scale.

- **Correlation Analysis**

  After the conversion of categorical variables to dummy variables, a correlation analysis can be conducted using Excel (Exhibit 4, Q2 tab) or R (R script file). The dependent variables are math,

writing, reading, and average test score (calculated by averaging all three subjects). The independent variables are all external factors in the data.

Discussion Point: Ask students what the purpose is of conducting a correlation analysis. This is done to determine whether any independent variables are highly correlated to each other, as this would cause collinearity and affect the significance of the regression analysis.

The correlation in both R and Excel will show that independent variables are not strongly correlated, which means there is no collinearity, and it is safe to proceed with regression.

- **Regression Analysis**

  Before beginning the regression analysis, ask students how many regressions they performed, what values they used to interpret the results of their regression and what methods can be used to validate the results of their regression analysis.

  In Excel, conduct four regression analyses, one for each subject and one on the average test scores (Exhibit 5, Q2 tab). The R-squared statistics across our regression analyses range from 0.22 to 0.32. This indicates that 24% of variance in math score, 22% of variance in reading score, 32% of variance in writing score, and 23% of variance in average testing scores can be explained by the independent variables used collectively. Although the R-square statistics are not very high, this is expected since the dependent variable is human academic performance and human behavior is difficult to predict. Therefore, the R-squared statistics are sufficient for our model.

  Next, students should plot the residuals to ensure the validity of the regression model (Exhibit 6, Q2 tab). The plots show no pattern and are randomly dispersed around the horizontal axis, which indicates the linear regression model is valid and can be used for analysis.

  In this case, the null hypothesis is the popular opinion that the educational system is a level playing field and the alternative hypothesis is that disparity caused by factors such as race exists in the system. The p-values of all independent variables are less than 0.05, which provides us with the evidence necessary to reject the null hypothesis and prove that correlation exists amongst our independent and dependent variables. Additionally, the significance F value is very small indicating that the probability of accepting the null hypothesis is very low within our regression model.

To further discuss the results of the regression analysis and conclude the extent to which each independent factor impacts student test scores, we can use R to run a regression analysis for each variable and subject. The results of the analysis using R and key points for discussion are as outlined:

- *Regression Results for Gender*

  Running a regression on math, reading, and writing scores with gender as the independent variable indicated that approximately 3-6% of the variance amongst test scores can be explained by gender (lines 359-362, 406-409, 453-456 in R script). On average, female students are positively correlated with average score and the score tends to increase by 3.9% if the student is female. In addition, male students tend to outperform female students in math by 4% but trail behind females in reading and writing by 7-10%. This point is further emphasized by the boxplots and density plots (Exhibit 7, Q2 tab).

- *Regression Results for Ethnicity*
  Running a regression on math, reading, and writing scores with ethnicity as the independent variable indicated that 2-5% of the variance amongst test scores can be explained by ethnicity (lines 368-371, 415-418, 462-465 in R script file). For students of race A, B, C, and D, the average performance would decrease by 9.9%, 7.8%, 6.0%, and 3.0% respectively, in comparison to race E. From the graphs in Exhibit 8, Q2 tab, it is apparent that those belonging to the ethnicity of group A needed extra help across all three subjects as their average test scores were the lowest. For the test scores of the remaining ethnic groups, group B was slightly below average, group C was average, group D was above average, and Group E consistently outperformed all other ethnicities.

- *Regression Results for Parental Education Level*
  Running a regression on math, reading, and writing scores with the variable of parental levels of education indicated that approximately 3-6% of the variance amongst test scores across all three subjects can be explained by the students' parent's level of education (lines 351-353, 398-400, 445-447 in R script file). If parents did not complete high school, the student's average performance decreased by 4.3%. However, if parents have a post-secondary degree, the student's performance increased by 4.4%. Reviewing the density plots and boxplots in Exhibit 9, Q2 tab also reveals similar insights.

- *Regression Results for Lunch Options*
  Running a regression on math, reading, and writing scores with lunch status as the independent variable indicated that approximately 5-12% of the variance amongst scores across all three subjects can be explained by the student's lunch status (lines 387-389, 434-436,481-483 in R script file). There were two lunch statuses depending on the student's financial need; students who required financial assistance were given free or reduced lunches. The strong relationship here implies that students who are from a more financially secure background tend to perform better academically. If a student had a free or reduced lunch, the performance score would decrease by 8.7% compared to students who had a standard lunch. As evidenced by a higher R-squared value in math scores, a student's financial situation has a more discernable impact on math scores rather than reading or writing scores. This point is further emphasized by the boxplots and density plots in Exhibit 10, Q2 tab.

- *Regression Results for Test Preparation Course*
  Running a regression on math, reading, and writing scores with test preparatory completion as the independent variables indicates that approximately 3-10% of the variance amongst test scores across all three subjects can be explained by the completion of test preparation courses (Lines 378-380,425-427, 472-474 in R script file). If a student has completed a test preparation course, their average academic performance would increase by 6.9% compared to student who had not completed test prep. Based on the R-squared values and the figures in Exhibit 11, Q2 tab, it seems apparent that students who took these extra preparation courses benefitted greatly. These students received higher average scores by 5% in math, 8% reading, and 11% in writing.

- *Impact of Multivariate Factors on Student Scores*
  Students may also run multivariate factors analysis using R to analyze the impact of multiple variables on test scores. This is a fascinating and useful tool that students can analyze for cases

further down the line. From Exhibit 12, Q2 tab, some interesting patterns can be observed. Firstly, improvement in scores is nearly unanimous for students who take the test preparation course regardless of their ethnicity or parental education history. This may indicate that no ethnicity inherently performs worse than others since scores increased for all those who took the test preparation course regardless of ethnicity. These patterns are amplified for reading and writing scores. For reading and writing scores, the parent's highest education level also seemed to highly impact their child's scores. Despite this, students whose parents have a bachelor's or master's degree still see an improvement in grades when they take the test preparation course. All three plots seem to indicate getting more students involved in test preparation courses regardless of their economic backgrounds or ethnicity would be beneficial. Although students who come from a more affluent socio-economic background tend to perform better than the rest of their cohort, this may be caused by their ability to have access to expensive tutors and purchasing the test preparation courses.

Assignment Question 3

How does the preparation course make an impact on scores of students in need? If student X (whose score is 65%) took the prep. course, what is the probability they will obtain at least an 80% on the next test?

Analysis

Conducting a conditional probability analysis on test scores for students with parents who have limited education or for students from non-affluent families can help drive insights on the impact of preparation courses for students in need.

First, we can calculate the impact of test preparation courses on students whose parents hold below-high-school degrees. Of the 178 students whose parents hold below-high-school degrees, only 57 students took test preparation courses. From Exhibit 13, Q3 tab, the probabilities of scoring above 80 in three subjects were higher for those who had taken preparation courses than those who have not taken them. In particular, there was a 6.7% average increase in probability of scoring above 80 in writing and reading while the increase in math was 3.5% after taking preparation courses.

Another observed relationship was the impact of preparation courses on non-affluent students via the proxy of reduced/standard lunch. From the plots in Exhibit 14, Q3 tab, test preparation had a larger effect on writing scores than math and reading scores. The probability of non-affluent students obtaining a 90% in writing increased around 2% after taking test preparation courses. It should be noted that non-affluent students have a low preparation course participation rate of 35%, although it is unclear whether this is due to lack of funding or motivation.

Discussion Points:

Based on this analysis, test preparation courses seem to have a positive effect on students in need, although this improvement is larger in linguistics than it is in math. This indicates that if more students in need were enrolled in test preparatory courses, this would positively influence their academic performance.

Students can also use conditional probability analysis to calculate the change in the probability of achieving a high-test score between students given that they take or do not take the test preparatory courses. Another method is using R to draw density plots for scores of students in need.

Determining Impact of Test Preparation Courses on a Student

In the given problem, we want to determine the impact of test preparation courses for a student who scored 65%. Students can first use the NORM.DIST() function using the data from the conditional probability analysis to determine the percentile a 65% scoring student is at. Students can then use the NORM.INV() function to calculate the equivalent score at the same percentile using the data from students that received test preparation.

The calculations yield a 6% increase from 65% to 71% after receiving test preparation courses. Although the student is unlikely to obtain a test score above 80% following the test preparation, this 6% improvement highlights the significant impact of providing additional resources to students (Exhibit 15, Q3 tab).

Assignment Question 4

Professor Kwan wants to consider all factors affecting student X's math scores. Student X is a male belonging to ethnic group B, has a modified lunch plan and their parents highest level of education was a bachelor's degree. Predict Student X's math score both before and after taking the prep. course.

After determining the effects of the preparation course based on the student X's previous grade alone, students now turn to predict a score given specific characteristics of a student, which can be done using their regression outputs from question 2. Using a SUMPRODUCT function for the coefficients for each factor and the assignment of 1 or 0 for the student, the using SUM to calculate the final value of these products, students can predict student X's score both before and after the preparation course. As seen in Exhibit 16, Q4 tab, we can predict that student X had a score of about 61% before and 65% after the preparation course.

Discussion: Comparing the findings from this question and the previous, students can see that conditional probability can be a helpful tool for generalities, while regression has stronger prediction power when specifics for the situation at hand are known. We saw that the conditional probability analysis gave us an answer in approximately the same range as the regression prediction, however the regression narrows this range and is generally more precise.

Assignment Question 5

If you were Professor Kwan, would you change anything in your classroom/teaching based on your findings? If so, why and how would this help further affirmative action in education?

Based on the analysis completed and insights uncovered, students are likely to recommend several changes to Professor Kwan as well as changes for educational institutions in general. Presented below is a non-exhaustive list of ideas that students may touch upon.

- A solution for improving male student scores could be mandatory daily in-class reading and daily journaling to ensure that all students are reading and writing at the required level.

- For students to improve their grades in math, a solution could lie in expanded opportunities for students to be tutored in areas in which they need more guidance and help.
- To improve overall academic performance, the school should increase the number of students who are participating in test preparatory courses. One way to increase participation could be to incentivize students with bonus marks for completing the test preparation courses for all three subjects. If students are unable to afford these courses the school may also want to partially subsidize these resources for students in need.
- Some solutions to improve the discrepancy in student scores amongst different ethnic groups could be to increase collaboration amongst students, creating diverse groups for assignments. This could benefit group A by having Group E members as a positive influence and provide a teaching experience for Group E as well. Additionally, government programs and advocacy groups should focus their resources for academic support on ethnic groups A and B. Lastly, there is a chance that the teachers themselves might be holding subconscious prejudices favoring one group and disparaging the other. To limit any bias in the future, tests can be marked blindly without looking at the student's name which could resolve this issue.
- Suggestions for alleviating the issue of low-income students performing poorly academically could be free after-school tutoring sessions to help students in need based on family income and/or academic performance. Although more affluent students tend to perform better than their non-affluent counterparts, this may be caused by their access to expensive tutors and test preparation courses. To alleviate this disparity, more resources can be put forth by the school such as subsidizing the test preparation courses and providing tutoring sessions to children of low-income families.
- Additionally, the school could start a peer-mentor program where top students are tutors and peer mentors for students whose are struggling academically.
- To alleviate the disparity amongst students whose parents have different educational histories, teachers can also provide additional explanations or readings to students whose parents did not complete post-secondary education.

CONCLUSION

Through their analysis of student scores, students will gain valuable single and multiple regression skillsets. To conclude the class, explain to students that although we cannot provide a "what happened" overview in this case, the scenario is a great way to get students thinking about issues bigger than themselves that are present in our society. Fostering a view of inclusiveness is an important trait in managers and this case is a great example of how to use analytics to further a social cause.

REFERENCES

1. School Nutrition Association. School Meal Trends and Stats. Retrieved December 14, 2020, from https://schoolnutrition.org/aboutschoolmeals/schoolmealtrendsstats/