CS555 Term Project

Lang Xue

U80614812

1. Research Scenario Description

A lot of researches have been done to explore the potential factors to affect the academic performance. The possible factors could be race, age, gender, family education levels, financial status and so on. Allen (1991) studied how race and gender affect the academic performance in US higher education. Dayioğlu (2007) studies the gender difference in academic performance in a large public university in Tukey. Adekitan (2020) used regression method to compare the academic performance between gender groups.

In this project, we would like to know whether the academic performance are affected by the demographic characteristics. In particular, we would like to know whether the academic performance is different across male and female. Besides, does this effect change when we control for the age.

1. Describe the data set

The data set records the student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features. It was collected by using school reports and questionnaires. This data set is about the performance in Portuguese language. The details can be found in <https://archive.ics.uci.edu/ml/datasets/Student+Performance>.

The dataset includes 649 observations. We focus on the school Gabriel Pereira. The data set is reduced to 423 observations. One observation has 0 exam score. We consider it as an outlier. It is been removed. The final data set contains 422 observations.

Among all the attributes, we select only 3 columns. The first column is the first period grade, range from 0 to 20. The second column is the sex, F for female and M for male. The last column is the student’s age.

The link to the data set is <https://www.kaggle.com/larsen0966/student-performance-data-set>.

1. Research Question

We would like to know whether the average grade is the same across male and female. Besides, does this effect change when we control for the age.

1. R Code Solution

|  |
| --- |
| student <- read.csv("student-por.csv")  table(student$school)  # only consider school = GP, it reduces observations from 649 to 423  mydata <- student[student$school == "GP", c("G1", "sex", "age")]  # remove outlier, the one with 0 score, sample size becomes n=422  mydata\_clean = mydata[mydata$G1 != 0,]  dim(mydata\_clean) # 422, 3  names(mydata\_clean)[1] = "score"  #write.csv(mydata\_clean, "mydata\_clean.csv")  # check distribution of the score (roughly normal)  par(mfrow=c(1,2))  with(mydata\_clean, hist(score))  with(mydata\_clean, qqnorm(score))  with(mydata\_clean, qqline(score))  # scatterplot  library(ggplot2)  ggplot(mydata\_clean, aes(x = age, y = score, col = sex)) + geom\_point() +  ggtitle("Score against age by gender")    # two sample t test  with(mydata\_clean, t.test(score ~ sex))  # test the score difference between male and female, control for age  fit1 = lm(score ~ sex + age, data = mydata\_clean)  summary(fit1) |

1. Execute your R code, Copy and Paste results here in this Box

|  |
| --- |
| Chart, scatter chart  Description automatically generated  Chart, histogram  Description automatically generated  > # two sample t test  > with(mydata\_clean, t.test(score ~ sex))  Welch Two Sample t-test  data: score by sex  t = 3.1939, df = 375.78, p-value = 0.001522  alternative hypothesis: true difference in means is not equal to 0  95 percent confidence interval:  0.2832086 1.1904564  sample estimates:  mean in group F mean in group M  12.33898 11.60215  >  > fit1 = lm(score ~ sex + age, data = mydata\_clean)  > summary(fit1)  Call:  lm(formula = score ~ sex + age, data = mydata\_clean)  Residuals:  Min 1Q Median 3Q Max  -5.2621 -1.6862 -0.0558 1.5202 6.8081  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 17.15762 1.51637 11.315 < 2e-16 \*\*\*  sexM -0.78227 0.22581 -3.464 0.000587 \*\*\*  age -0.28797 0.09018 -3.193 0.001513 \*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 2.298 on 419 degrees of freedom  Multiple R-squared: 0.04748, Adjusted R-squared: 0.04294  F-statistic: 10.44 on 2 and 419 DF, p-value: 3.748e-05 |

1. Conclusion

The average grade is different across male and female since the p-value in the two sample t-test is below 5% level of significance. Controlling for student’s age, average grade is different across male and female since the p-value for the coefficient of sex in the regression model is below 5% level of significance

Reference:

1. Allen, Walter R., and Nesha Z. Haniff. "Race, gender, and academic performance in US higher education." College in Black and White: African American students in predominantly White and in historically Black public universities (1991): 95-109.
2. Dayioğlu, Meltem, and Serap Türüt-Aşik. "Gender differences in academic performance in a large public university in Turkey." Higher Education 53.2 (2007): 255-277.
3. Adekitan, Aderibigbe Israel, and Olamilekan Shobayo. "Gender-based comparison of students’ academic performance using regression models." Engineering and Applied Science Research 47.3 (2020): 241-248.