Exploring Education Year and Correlated Factors in 1994 America

Yuwei (Johnny) Meng - 1007824810

2023 Winter JSC270S

Professor Carolina Nobre

14 February 2023

Introduction

Demographics have long been an interesting set of data to sociologists because they reveal social patterns. Particularly, in the UC Irvine Machine Learning Repository, there is a set of demographic data, extracted from the 1994 Census database. This extraction results in a data frame of dimension 32561×15, including both categorical and numerical attributes such as age, sex, education, race, and more about people living in a region of the U.S. in 1994. This dataset was originally created for predicting whether a person makes over 50K a year based on their demographic characteristics. Nevertheless, more insights can be unearthed from this comprehensive dataset.

Besides income, education level is often of interest to researchers as well. One reason is that education can detect poverty level, social inequality, and more. Thus, using this old dataset, the current study aims to build a model to predict people's education level based on other demographic factors and then attempts to reveal social patterns in the U.S. in 1994 using this model.

Methodology

The selected response variable for this study is *education_num*, which represents the year of education a person has taken. The selected independent variables include *sex*, *race*, and *age*. *Sex* and *race* are chosen for this analysis assuming that inequality was present in the U.S. society in 1994. The *age* variable is selected considering that older people should have spent longer time on study than younger people. Further, since 95% of the subjects are white or black people (just using the term in the data; no offense to colored people), this study decides to exclude observations with other races to make the model more easily interpretable.

Results & Discussion

Before fitting the model, a histogram was plotted for the response variable. The graph shows the peak at 9 years, with more people having more than 9 years of education and fewer people having less than 9 years. This rough symmetry satisfies a necessary assumption for linear regression. Then, a linear model is built. The fitted equation is as follows:

$$\text{education\_num} = 9.2346 - 0.0227(\text{male}) + 0.6463(\text{white}) + 0.0070(\text{age})$$

Several discoveries can be concluded from this linear model. Firstly, the coefficient for male is 0.0227, with a corresponding p-value of 0.465, which is not statistically significant. This manifests that at the same age and with the same race, a man and a woman in 1994 were expected to attain the same year of education on average. This discovery shows that men and women were relatively equal in education in 1994, which should be appreciated. Secondly, the coefficient for white is 0.6463 with a p-value smaller than 0.001. This reveals that on average, white people were achieving higher education than colored people in 1994, which was an issue because everyone should have equal rights to education regardless of race. Lastly, the coefficient for age is 0.0070 with a p-value of less than 0.001. This confirms with the earlier assumption that older people attained higher education than younger people.

Nevertheless, there are some limitations to this linear model. The major one is that the R^2 value is only 0.007, or only 0.7% of the variability is explained by this model, which is a poor performance. Therefore, the above linear relationship between the chosen variables should be further examined with more data. On the other hand, this poor performance might overthrow this model, leading to concluding that it is possible that education was more attainable in 1994 than expected, which is a good sign of an equal society. Regardless of the validity, it is important that sociologists and educators strive to provide accessible education to everyone, no matter what race, sex, or nation they belong to.