The correlation between work hours per week and other variables for people from 1994

1 Background

The data set used in this assignment was extracted from the U.S. census bureau database of 1994, and contains information of people such as their age, education, capital gain, work hours per week, etc. There are a total of 32561 data points in this data set, or 30162 if we remove data points with unknown attributes. Moreover, each data point in the data set has a `fnlwgt` attribute that denotes the its weight, i.e., the number of people the data point represent.

2 Motivation

It seems reasonable that people's work hours per week is associated with other factors such as their education background or which gross income group they are in. In the last question of this assignment, we try to fit a linear regression model that should be capable to predict how many hours do people from 1994 work per week using other variables from our data set (`sex`, `education_num`, and `gross_income_group` in this case).

3 Methods

It follows from oour motivation that we can try to fit a linear regression model with `hours_per_week` as the dependent variable (the one that we are interested in predicting) and `sex`, `education_num`, and `gross_income_group` as the independent variable. The exact method that was being used is OLS(Ordinary Least Squares) from the package `statsmodels`, which fits a linear regression model that minimises the sum of square differences between the observations and predictions. The resulting model should give four coefficents from the linear regression formula:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3,$$

where Y denotes work hours per week, X_1 denotes sex, X_2 denotes education length in years, and X_3 denotes gross income group.

4 Results

The linear regression model from the OLS algorithm has

$$R^2 = 0.094,$$

which is reasonably good considering that we are only using three independet variablels. In addition, the model provides us the coefficents for all three dependent variabels and the y-intercept:

$$\beta_0 = 31.42,\ \beta_1 = 5.10,\ \beta_2 = 0.45,\ \beta_3 = 4.52.$$

5 Conclusion

Overall, the results meet our expectation. According to the site *Striking Women*, much of women's work are irregular, home-based or within a family-run business in the 19th century[1], thus men tend to work more than women. This is consistent with the model as β_1 is positive, meaning that `hours_per_week` increases if we look at data for men instead of women. In addition, the model tells us that if we move from the "$<= 50,000$" income group to the "$> 50,000$ group, we should expect a 4.52 work hours per week increase. This is as expected because the length of work and total income should be positively associated. Lastly, the coefficent for the length of education in years is small and positive (0.45) which suggests that some work that requires higher education may also require occassionaly working overtime.

To conclude, sex, length of education in years, and gross income group are associated with work hours per week for people from 1994. Each of the independent variables have positive coefficent. Further, one's sex and gross income group are the most significant.

[1] *Women and work in the 19th century*, (Striking Women).