Assignment 2 Part III

The dataset of interest, the UCI Census Income dataset, is a subset of data from the 1994 US nationwide census. It was created by Ronny Kohavi and Barry Becker to explore the relation between individual income and other factors and to predict whether an individual's income is greater than 50,000 USD. It contains data on 14 census attributes for 48842 individuals and includes both categorical and numerical data. Attributes include "age", "education", and "capital-gain".

For my linear regression model for this dataset, I am interested in answering: Do older individuals working in office jobs with greater annual income tend to have greater capital gains? My rationale for this question is that older individuals working in office jobs who have high incomes are likely to be more financially stable and experienced compared to younger individuals. If this is true, they would be more likely to successfully invest in assets such as real estate, stocks, and bonds, especially if these assets relate to their profession. This seems reasonable since many office jobs are in the real estate or financial sectors. This heuristic reasoning is thus the motivation for this question.

Since I have more than one variable to vary, I will use a multiple linear regression model. The response variable is *capital_gain* since I want to explore the change in capital gains for my question. The feature variables are *age* and *gross_income_group* since I want to vary the age and income of individuals for the change in the response. Since I am investigating individuals who have office-based occupations, I only select data from individuals whose values for *occupation* are 'tech-support', 'sales', 'exec-managerial', or 'adm-clerical'. This is because jobs like sales and management are frequently office-based, while other jobs such as 'farming-fishing and 'transport-moving' require more manual labor. In addition, I will consider the 74 people with capital gains of exactly 99,999 USD as outliers since the value is significantly greater than the capital gains in the rest of the data.

After fitting the model, my findings are as follows: the intercept, the estimated expected *capital_gain* for individuals of age 0 with income < 50K, is -245.52. Since often only adults can have

capital gains, this value does not have a real-world interpretation. The coefficient for *gross_income_group* is 1887.5118, which means that there is an estimated expected increase of 1887.5 USD for *capital_gain* of individuals with >= 50K income compared to individuals with < 50K income. The coefficient for *age* is 10.9223, which means that there is an estimated expected increase of 10.9 USD for *capital_gain* as *age* increases by one. The values all have p-values 0.000 or 0.001, indicating high or very high confidence that they are different from 0 (the alternative hypothesis). Finally, my model has an adjusted R-squared value of 0.100 and an RMSE of 7475785.7. Since R-squared should be as close to 1 and the RMSE as low as possible for a good fit, this indicates that my model is a weak fit of the data.

Overall, to answer my question, older individuals working in office jobs with greater annual income do tend to have greater capital gains. There is an increase in *capital_gains* as *age* increases and *gross_income_group* changes from < 50K to >= 50K. Note that the increase in *capital_gains* for the change in *gross_income_group* (1887.5 USD) is much larger than that for a one-unit increase in *age* (10.9 USD). This is likely because the data is more spread out for age, whereas there are only two gross income groups which may be very different in demographic composition.

The p-values of the coefficients indicate they are statistically significant and that we can be reasonably certain there is a general increase in the data as age and gross income increases. However, the low R-squared and high RMSE values indicate that our model is likely not a good fit of the data. This dichotomy exists likely because there are more parameters or covariates that influence the correlation, and they have not been controlled for in the model. For instance, regarding my initial heuristic reasoning for my question, it may be that the number of years in education, rather than age, better correlate with success with financial assets. But if older people have generally spent more time in education than younger people, this results in the weak correlation observed from my model. In addition, we assume that the expected value of the error term is zero for our regression model, which might not be the case here due to the same reason.