Assignment 2 Part II

Link to repository: https://github.com/ryancys1234/JSC270_HW2_2022_rshi

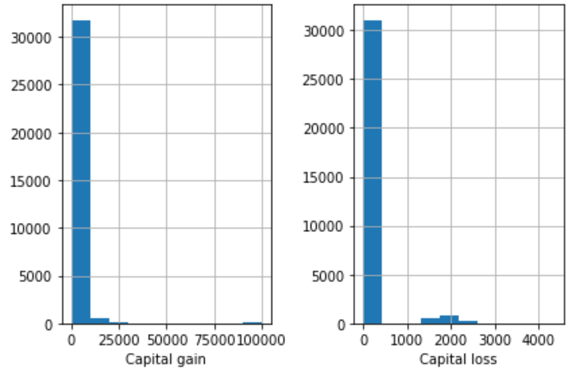
Initial Data Exploration:

1. The columns that are numerical data (int64 in the above table) are 'age', 'fnlwgt', 'education_num', 'capital_gain', 'capital_loss', and 'hours_per_week'. In the text file, these columns are denoted as continuous data.

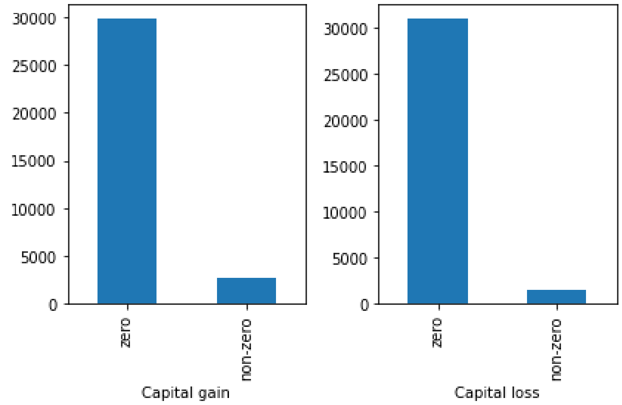
   All the other columns are categorical data. This is expected since variables like age, capital gain and loss, and hours per week are all described by numbers, whereas education, occupation, and race are better described by categories. Note that 'fnlwgt', the final weight, is indicated as continuous data in the text file description of the data.

2. Missing values are represented with ' ?'.

3.

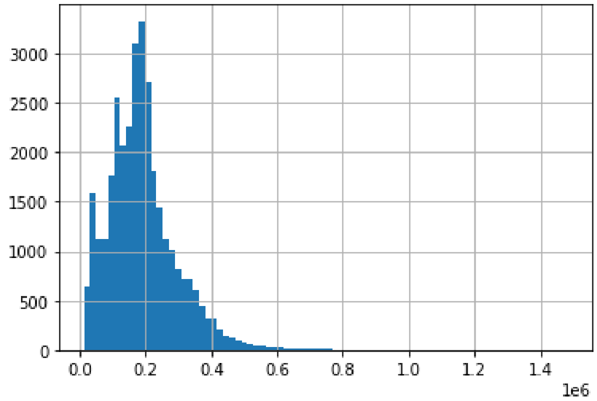


   I think these variables should be transformed to categorical variables. In the plots of capital gain and loss shown above, there seems to be many zero values in both capital gain and loss. As such, I will separate the zero values from the non-zero values in new categorical variables called “capital_gain_category” and “capital_loss_category”, which will have the values "zero" and "non-zero".

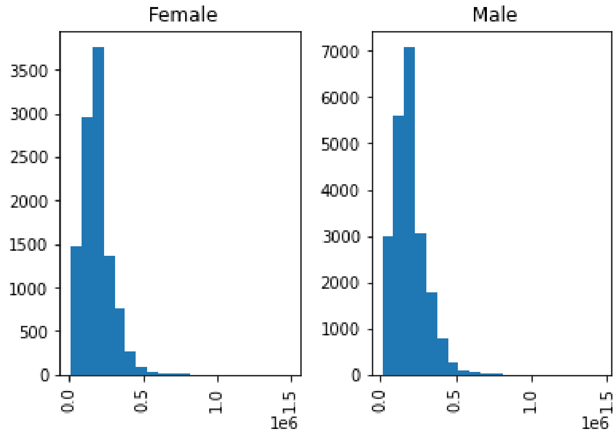


From the plots of “capital_gain_category” and “capital_loss_category” above, there appears to be significantly more zero values than non-zero values for the variables “capital_gain” and “capital_loss”.

4. Below is the histogram of the distribution of fnlwgt:



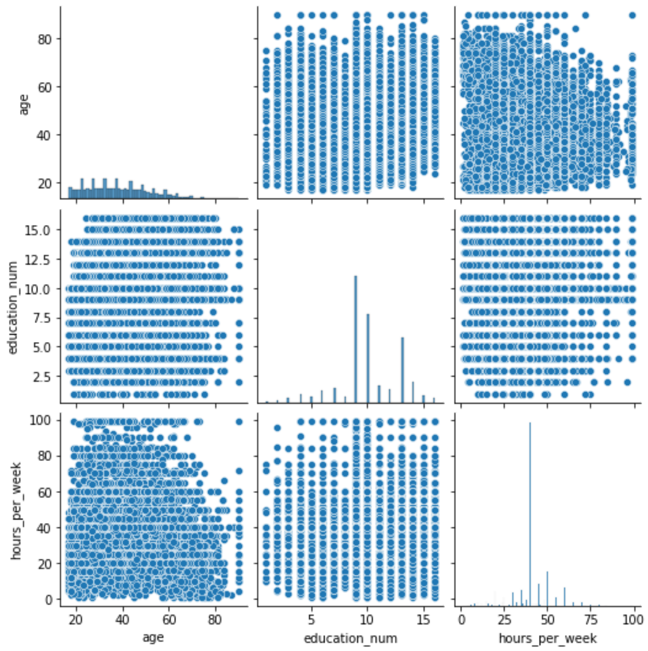
The variable fnlwgt is not symmetrically distributed; it appears to be right-skewed in the above graph. On the left side of the mode, there appears to be smaller modes, so the distribution is multimodal. On the right side, the distribution of values roughly follows the negative exponential graph.



The distribution of fnlwgt is identical between men and women, as shown in the above plots of fnlwgt for males and females. For instance, the individual distributions are both right-skewed. However, there is more data for fnlwgt for males than females. There does not appear to be any significant outliers outside of the tails, and so they will not be excluded.

Correlation

1. a) Below are the correlation plots for 'age', 'education_num', 'hours_per_week':



None of the variables appear to be correlated. Correlation between two variables is when adjusting the value of one variable leads to a change in the value of the other, and vice versa. As shown in the non-diagonal subplots, this does not occur since in each subplot, there is a variable whose value remains constant as the other variable varies. (This is shown by the vertical or horizontal lines in the subplots.)

b) The variable pairs with correlation coefficient > |0.1| are “capital_gain” and “education_num” (coefficient 0.122630), and “hours_per_week” and “education_num” (coefficient 0.148123).

The p-values for the Pearson test on both pairs are 0.000. This means we are very confident about the alternative hypothesis that the coefficients are different from 0. In other words, we are confident that the variable pairs are positively correlated. This is expected since it makes sense that the higher a person’s level of education is, the more likely they are to earn more and work hard (e.g., as determined by the number of hours they work). It could also be that someone who works more also earns more, like if they are paid by hourly wage, and this itself is correlated with their level of education.

c) For males, the coefficient is 0.060486 and the p-value is 0.000. For females, the coefficient is -0.017899 and the p-value is 0.063. Thus, the correlation between education_num and age is positive for males and negative for females. The coefficient for females has a higher absolute value, while the p-value for females (0.063) is also higher (above 0.05 but below 0.1), indicating less certainty of this correlation (the null hypothesis is that they are not correlated, while the alternative is that they are).

This means that we are more certain that education_num and age positively correlate for males than we are certain that they negatively correlate for females. This is somewhat expected, since older males are likely to have spent more years in education, such as by attaining master's or PhD degrees. Furthermore, notice that the dataset was created in 1994, a time when more women entered higher education compared to earlier generations. It might be possible that back then, there were many older women who were less educated than younger ones, thus creating the observed negative correlation.

d) The covariance matrix gives us the covariance between education_num and hours_per_week and the variance for each variable. The covariance between education_num and hours_per_week is 4.705338, which is positive and thus indicates that as the variables grow in the same direction (they either increase together or decrease together). The variance of education_num is 6.618890, while the variance of hours_per_week is 152.458995, which is significantly higher. This is likely because different people in different positions and jobs have different working environments, meaning they are likely to spend time differently, and also that some people would want to work longer for promotions or bonuses, for instance.

Regression

1. a) Yes. The intercept, or the value for the estimated expected hours_per_week for females, is 36.4104. The coefficient for males is 6.0177, which means that compared to females (the intercept), the expected hours_per_week for males is 6.0177 higher than that for females. Thus, with a higher expected hours_per_week, males tend to work more hours.

   b) The general trend remains the same in that males still tend to work more hours than females, but the difference between the expected value for males and females is smaller (5.9709 for this model, compared to 6.0177 for the previous). The p-value for the coefficient of education_num is 0.000, indicating that it is statistically significant. The 95% confidence interval for the coefficient of education_num is [0.647, 0.748], while the coefficient itself is 0.6975.

   c) For the first model ('hours_per_week' ~ 'sex'), the coefficient of 'sex' is 6.0177. The interpretation for this is the same as in 1a).

   For the second model ('hours_per_week' ~ 'sex' and 'education_num' is the control), the coefficient of 'sex' for males (with education_num of 0) is 5.9709, with the data on females (with education_num of 0) as the intercept. This means that as we move from data on females to males who have not spent time in education, the expected number of hours worked in a week increases by 5.9709, indicating that such males tend to work 5.9709 more hours on average compared with such females.

   For the third model with 'sex', 'gross_income_group', and 'education_num', the coefficient of 'sex' is 5.1010 for males (with <= 50K for gross_income_group and education_num of 0). As before, the data on females (with <= 50K for gross_income_group and education_num of 0) is the intercept. This means that as we move from data on females to males whose gross income is below 50K and who have not spent time in education, the expected number of hours worked in a week increases by 5.1010, indicating that such males tend to work 5.1010 more hours on average compared with such females.

The statistics that can help to determine the ‘best’ model are the adjusted R-squared and RMSE values. The closer the R-squared value is to 1, the better the fit is, while RMSE values should ideally be as low as possible (i.e., near 0).

The adjusted R-squared for the first model with only ‘sex’ and ‘gross_income_group’ is 0.053, for the second model (‘sex’ + ‘education_num') is 0.074, and for the last model (‘sex’ + ‘education_num’ + ‘gross_income_group’) is 0.094. They are similar, but since the value for the last model is the closest to 1, it is the best fit. Similarly, the RMSE value for the first model is 144, for the second is 141, and for the last is 138. Again, they are similar, but since the RMSE value for the last is the lowest, it is the best fit.

Overall, the model with ‘sex’ + ‘education_num’ + ‘gross_income_group’ is the best fit.

Bonus question: on next page (handwritten)

Bonus question:

- The closed-form of $\hat{\beta}_1$, the estimate of the slope parameter, is $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$
  where $\bar{x}$ is the sample mean for variable x, $\bar{y}$ the sample mean for variable y.

- The definition of r_{xy}, the sample correlation coefficient for variables x and y, is
  $r_{xy} = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1) s_x s_y}$, where $\bar{x}, \bar{y}$ are as before, $s_x = \sqrt{s_x^2} \geq 0$ is the sample standard deviation of X, $s_y = \sqrt{s_y^2} \geq 0$ is the sample standard deviation of y.

- Notice $\dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)\, s_x^2} = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1) s_x s_y} \dfrac{s_y}{s_x} = r_{xy} \dfrac{s_y}{s_x}$

  where $s_x^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{(n-1)}$ is the sample variance for x,

- Thus, $\hat{\beta}_1 = r_{xy} \dfrac{s_y}{s_x}$ is the relation between $\hat{\beta}_1$ and r_{xy}.

- Overall, as the standard deviations of the dependent and independent variables become more similar, the rate of change in the dependent variable for a one-unit change in the independent variable becomes closer to the measure of how well they correlate.

  If the standard deviation of two variables are equal, the estimate of the slope parameter is equal to the sample correlation coefficient.

  If the standard deviation of the dependent variable is greater than that of the independent variable, the estimate for the slope parameter is larger than the sample correlation coefficient.

  If the standard deviation of the dependent variable is smaller than that of the independent variable, the estimate for the slope parameter is smaller than the sample correlation coefficient.

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.